

Fast Routing Table Construction Using Small Messages

Christoph Lenzen*

Dept. Computer Science & Applied Mathematics
Weizmann Institute of Science
Rehovot 76100, Israel

Boaz Patt-Shamir†

School of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, Israel

Abstract

We describe a distributed randomized algorithm computing approximate distances and routes that approximate shortest paths. Let n denote the number of nodes in the graph, and let HD denote the *hop diameter* of the graph, i.e., the diameter of the graph when all edges are considered to have unit weight. Given $0 < \varepsilon \leq 1/2$, our algorithm runs in $\tilde{\mathcal{O}}(n^{1/2+\varepsilon} + \text{HD})$ communication rounds using messages of $\mathcal{O}(\log n)$ bits and guarantees a stretch of $\mathcal{O}(\varepsilon^{-1} \log \varepsilon^{-1})$ with high probability. This is the first distributed algorithm approximating weighted shortest paths that uses small messages and runs in $\tilde{o}(n)$ time (in graphs where $\text{HD} \in \tilde{o}(n)$). The time complexity nearly matches the lower bounds of $\tilde{\Omega}(\sqrt{n} + \text{HD})$ in the small-messages model that hold for *stateless* routing (where routing decisions do not depend on the traversed path) as well as approximation of the weighted diameter. Our scheme replaces the original identifiers of the nodes by labels of size $\mathcal{O}(\log \varepsilon^{-1} \log n)$. We show that no algorithm that keeps the original identifiers and runs for $\tilde{o}(n)$ rounds can achieve a polylogarithmic approximation ratio.

Variations of our techniques yield a number of fast distributed approximation algorithms solving related problems using small messages. Specifically, we present algorithms that run in $\tilde{\mathcal{O}}(n^{1/2+\varepsilon} + \text{HD})$ rounds for a given $0 < \varepsilon \leq 1/2$, and solve, with high probability, the following problems:

- $\mathcal{O}(\varepsilon^{-1})$ -approximation for the Generalized Steiner Forest (the running time in this case has an additive $\tilde{\mathcal{O}}(t^{1+2\varepsilon})$ term, where t is the number of terminals);
- $\mathcal{O}(\varepsilon^{-2})$ -approximation of weighted distances, using node labels of size $\mathcal{O}(\varepsilon^{-1} \log n)$ and $\tilde{\mathcal{O}}(n^\varepsilon)$ bits of memory per node;
- $\mathcal{O}(\varepsilon^{-1})$ -approximation of the weighted diameter;
- $\mathcal{O}(\varepsilon^{-3})$ -approximate shortest paths using the labels $1, \dots, n$.

*Supported by the Swiss Society of Friends of the Weizmann Institute of Science and by the Swiss National Science Foundation (SNSF).

†Supported in part by the Israel Science Foundation (grant 1372/09) and by Israel Ministry of Science and Technology.

1 Introduction

Constructing routing tables is a central task in network operation, the Internet being a prime example. Besides being an end goal on its own (facilitating the transmission of information from a sender to a receiver), efficient routing and distance approximation are critical ingredients in a myriad of other distributed applications.

At the heart of any routing protocol lies the computation of short paths in weighted graphs, where edge weights may reflect properties such as link cost, delay, bandwidth, reliability etc. In the distributed setting, an additional challenge is that the graph whose shortest paths are to be computed serves also as the platform carrying communication between the computing nodes. The result of this double role is an intriguing interplay between two metrics: the given shortest paths metric and the “natural” communication metric of the distributed system. The first metric is used for the definition of shortest paths, where an edge weight represents its contribution to path lengths; the other metric is implicit, controlling the time complexity of the distributed computation: each edge is tagged by the time it takes a message to cross it. If these two metrics happen to be identical, then computing weighted shortest paths to a single destination is trivial (for the all-pairs problem, see below). For the general case, the standard normalization is that messages cross each link in unit time, regardless of the link weight; this assumption is motivated by network synchronization. On the other hand, the length of the message must be taken into account as well. More precisely, in the commonly-accepted **CONGEST** model of network algorithms [21], it is assumed that all link latencies are one unit and messages have fixed size, typically $\mathcal{O}(\log n)$ bits, where n denotes the number of nodes.

The classical algorithm for computing shortest path distributively is the distributed variant of the Bellman-Ford algorithm. This algorithm is used in many networks, ranging from local to wide area networks. The Bellman-Ford algorithm enjoys many properties that make it an excellent distributed algorithm (locality, simplicity, self-stabilization). However, in weighted graphs, its time complexity, i.e., the number of parallel iterations, may be as high as $\Omega(n)$ for a single destination. This is in sharp contrast with the $\mathcal{O}(\text{HD})$ time needed to compute *unweighted* shortest paths to a single destination, where HD denotes the unweighted “hop-diameter” of the network. The difference between n and HD can be huge; suffices to say that the hop-diameter of the Internet is estimated to be smaller than 50. Intuitively, the problem originates in the fact that the Bellman-Ford algorithm explores paths in a hop-by-hop fashion, and the aforementioned superposition of metrics may result in a path that is weight-wise short, but consists of $\Omega(n)$ edges. If shortest paths have at most $\text{SPD} \in \mathbb{N}$ edges, then it suffices to run the Bellman-Ford algorithm for SPD communication rounds. Indeed, the running time of a few distributed algorithms is stated as a function of this or a similar parameter for exactly this reason (e.g., [6, 15, 16]).

To the best of our knowledge, no distributed algorithm for computing (approximate) weighted shortest paths in $o(\text{SPD})$ time in the **CONGEST** model was known to date. In this paper we present a distributed algorithm that computes approximate all-pairs shortest paths and distances using small messages, in time that nearly matches the lower bound of $\tilde{\Omega}(\sqrt{n} + \text{HD})$.

1.1 Detailed Contributions

Our main technical contribution, presented in [Section 4](#), is an algorithm that, using messages of size $\mathcal{O}(\log n)$, constructs, for any $0 < \varepsilon \leq 1/2$, in $\tilde{\mathcal{O}}(n^{1/2+\varepsilon} + \text{HD})$ rounds node labels¹ of size $\mathcal{O}(\log \varepsilon^{-1} \log n)$ and routing tables of size $\tilde{\mathcal{O}}(n^{1/2+\varepsilon})$ facilitating routing and distance estimation with stretch $\mathcal{O}(\varepsilon^{-1} \log \varepsilon^{-1})$. We show that assigning new labels to the nodes is unavoidable by proving that any (randomized) algorithm achieving polylogarithmic (expected) stretch without relabeling must run for $\tilde{\Omega}(n)$ rounds. The running time of our algorithm is close to optimal, since known results [\[7, 8, 23\]](#) imply that computing such an approximation in the **CONGEST** model must take $\tilde{\Omega}(\sqrt{n} + \text{HD})$ rounds.

Our algorithm comprises two sub-algorithms that we believe to be of interest in their own right. One is used for short-range routing (roughly, for the closest \sqrt{n} nodes), the other for longer distances. The short-range algorithm constructs a hierarchy in the spirit of Thorup-Zwick distance oracles [\[30\]](#): A recursive structure of uniformly sampled “landmarks” is used to iteratively reduce the number of routing destinations (and routes) that need to be learned, and repeated use of the triangle inequality shows that the stretch is linear in the number of recursion stages. While this idea is not new, our main challenge is to implement the algorithm using small messages; to this end, we introduce a bootstrapping technique that, combined with a restricted variant of Bellmann-Ford (that bounds the hop range and the number of tracked sources), allows us to construct low-stretch routing tables for nearby nodes.

This approach runs out of steam (i.e., exceeds our target complexity) beyond the closest $\mathcal{O}(\sqrt{n})$ nodes, so at that point we switch to the “long-distance” scheme. The basic idea in this scheme is to pick roughly \sqrt{n} random nodes we call the *skeleton* nodes, and to compute all-to-all routing tables for them. This is achieved by simulating the spanner construction algorithm by Baswana and Sen [\[3\]](#). Again, the crux of the matter is an efficient implementation of this approach using small messages. To this end, we first construct a spanner of a graph defined by the skeleton nodes and shortest paths between them. Due to the small number of skeleton nodes and the reduced number of edges (thanks to the spanner construction), we can afford to broadcast the entire skeleton-spanner graph, thereby making skeleton routing information common knowledge. In addition, we can mark the corresponding paths in the original graph quickly. Here too, our main low-level tool is the restricted Bellmann-Ford algorithm that bounds both the range and the load.

Using variants of our techniques, in [Section 5](#) we derive efficient solutions to several related problems (all statements hold with high probability).

- For the Generalized Steiner Forest (GSF) problem we obtain, for any $0 < \varepsilon \leq 1/2$, an $\mathcal{O}(\varepsilon^{-1})$ -approximation within $\tilde{\mathcal{O}}((\sqrt{n} + t)^{1+\varepsilon} + \text{HD})$ rounds, where t is the number of terminals. This should be contrasted with the best known distributed approximation algorithm for GSF [\[15\]](#), which provides $\mathcal{O}(\log n)$ -approximation in time $\tilde{\mathcal{O}}(\text{SPD} \cdot k)$, where SPD is the “shortest paths diameter,” namely the maximal number of hops in any shortest path, and k is the number of terminal components in the GSF instance.
- For any $k \in \mathbb{N}$, we obtain an $\tilde{\mathcal{O}}((\sqrt{n})^{1+1/k} + \text{HD})$ -time algorithm that constructs labels of size

¹We remark that our use of the term differs from the common definition in that we distinguish between the auxiliary routing information stored by the nodes (the tables) and the (preferably very small) labels replacing the original node identifiers as routing address.

$\mathcal{O}(k \log n)$ and local tables of size $\tilde{\mathcal{O}}(n^{1/(2k)})$, and produces distance estimations with stretch $\mathcal{O}(k^2)$. Compare with the recent distributed algorithm [6] that attains the same local space consumption at running time $\tilde{\mathcal{O}}(\text{SPD} \cdot n^{1/(2k)})$ and stretch $4k - 1$.

- Given any $0 < \varepsilon \leq 1/2$, we can compute an $\mathcal{O}(\varepsilon^{-1})$ -approximation of the diameter within $\tilde{\mathcal{O}}(n^{1/2+\varepsilon} + \text{HD})$ rounds. We show that the standard construction yielding a lower bound $\tilde{\Omega}(\sqrt{n} + \text{HD})$ extends to this problem, implying that also for this special case our solution is close to optimal.
- Employing a different routing mechanism for the short-range scheme, we can assign the fixed labels of $1, \dots, n$. This comes at the expense of a larger stretch of $\mathcal{O}(\varepsilon^{-3})$ within $\tilde{\mathcal{O}}(n^{1/2+\varepsilon} + \text{HD})$ rounds, for any $0 < \varepsilon \leq 1/2$.

1.2 Related Work

There are many centralized algorithms for constructing routing tables; in these algorithms the goal is usually to minimize space without affecting the quality of the routes too badly. We briefly discuss them later, since our focus is the distributed model. At this point let us just comment that a naïve implementation of a centralized algorithm in the **CONGEST** model requires $\Omega(|E|)$ time in the worst case, since the whole network topology has to be collected at a single node just for computation.

Practical distributed routing table construction algorithms are usually categorized as either “distance vector” or “link state” algorithm (see, e.g., [26]). Distance-vector algorithms are variants of the Bellman-Ford algorithm [4, 11], whose worst-case time complexity in the **CONGEST** model is $\Theta(n^2)$. In link-state algorithms [19, 20], each routing node collects the complete graph topology and then solves the single-source shortest path problem locally. This approach has $\Theta(|E|)$ time complexity. While none of these algorithms uses relabeling, it should be noted that the Internet architecture in fact employs relabeling (IP addresses, which are used instead of physical addresses, encode some routing information).

From the theoretical perspective, as mentioned above, there has not been much progress in computing weighted shortest paths beyond the “shortest path diameter” (we denote by SPD) even for the single-source case: see, e.g., [6] and references therein. For the unweighted case, an $\mathcal{O}(n)$ -time algorithm for exact all-pairs shortest-paths was recently discovered (independently) in [14] and [22]. These algorithms do not relabel the nodes. In addition, a randomized $(3/2)$ -approximation of HD is given in [22], and a deterministic $(1 + \varepsilon)$ -approximation is provided by [14]. Combining results, [14] and [22] report a randomized $(3/2)$ -approximation of the unweighted diameter in time $\tilde{\mathcal{O}}(n^{3/4})$.

In [7], a lower bound of $\tilde{\Omega}(\sqrt{n})$ on the time to construct a shortest-paths tree of weight within a $\text{poly}(n)$ of the optimum is shown; this immediately implies the same lower bound on routing (more precisely, on *stateless* routing, where routing decisions depend only on the destination and not on the traversed path). To the best of our knowledge, the literature does not state any further explicit lower bounds on the running time of approximate shortest paths or distance estimation algorithms, but a lower bound of $\tilde{\Omega}(\sqrt{n})$ can be easily derived using the technique used in [7] (which in turn is based on [23]). In [12] it is shown that in the **CONGEST** model, approximating the diameter of unweighted graphs to within a factor of $3/2 - \varepsilon$ requires $\tilde{\Omega}(\sqrt{n})$ rounds. For the unweighted case, we extend this result to arbitrary approximation ratios.

In the Generalized Steiner Forest problem (GSF), the input consists of a weighted graph and a set of *terminal nodes* which is partitioned into subsets called *terminal components*. The task is to find a set of edges of minimum weight so that the terminal components are connected. Historically, the important special case of a minimum spanning tree (all nodes are terminals, single terminal component) has been the target of extensive research in distributed computation. It is known that in the **CONGEST** model, the time complexity of computing (or approximating) an MST is $\tilde{\Omega}(\sqrt{n} + \text{HD})$ [7, 8, 23]. This bound is essentially matched by an exact deterministic solution [13, 17]. An $\mathcal{O}(\log n)$ -approximate MST is presented in [16], whose running time is $\mathcal{O}(\text{SPD})$, where SPD is the “shortest path diameter” mentioned previously. For the special case of Steiner trees (arbitrary terminals, single component), [5] presents a 2-approximation algorithm whose time complexity is $\tilde{\mathcal{O}}(n)$ (which can easily be refined to $\tilde{\mathcal{O}}(\text{SPD})$). For the general case, [15] presents an $\mathcal{O}(\log n)$ -approximation algorithm whose time complexity is $\tilde{\mathcal{O}}(\kappa \cdot \text{SPD})$, where κ is the number of terminal components.²

We now turn to a very brief overview of centralized algorithms. Thorup and Zwick [29] presented an algorithm that achieves, for any $k \in \mathbb{N}$, routes of stretch $2k - 1$ using $\tilde{\mathcal{O}}(n^{1/k})$ memory. In terms of memory consumption, it has been established that this scheme is optimal up to a constant factor in worst-case stretch w.r.t. routing [24]. This result has been extended to the average stretch, and tightened to be exact up to polylogarithmic factors in memory for the worst-case stretch [1]. For distance approximation, the Thorup-Zwick scheme is known to be optimal for $k = 1, 2, 3, 5$ and conjectured to be optimal for all k (see [31] and references). The algorithm requires relabeling with labels of size $\mathcal{O}(k \log n)$. It is unclear whether stronger lower bounds apply to name-independent routing schemes (which keep the original node identifiers); however, for $k = 1$, i.e., exact routing, trivially $\mathcal{O}(n \log n)$ bits suffice (assuming $\mathcal{O}(\log n)$ -bit identifiers), and Abraham et al. [2] prove a matching upper bound of $\tilde{\mathcal{O}}(\sqrt{n})$ bits for $k = 2$.

A closely related concept is that of *sparse spanners*, introduced by Peleg and Schäffer [25]. A k -spanner of a graph is obtained by deleting edges, without increasing the distances by more than factor k . Similarly to compact routing tables, it is known that a $(2k - 1)$ -spanner must have $\tilde{\Omega}(n^{1+1/k})$ edges for some values of k , this is conjectured to hold for all $k \in \mathbb{N}$, and a matching upper bound is obtained by the Thorup-Zwick construction [30]. If an additive term in the distance approximation is permitted, the multiplicative factor can be brought arbitrarily close to 1 [9]. In contrast to routing and distance approximation, there are extremely fast distributed algorithms constructing sparse spanners. Our long-range construction rests on an elegant algorithm by Baswana and Sen [3] that achieves stretch $2k - 1$ vs. $\mathcal{O}(n^{1+1/k})$ expected edges within $\mathcal{O}(k)$ rounds in the **CONGEST** model.

2 Model

In this section we define the model of computation and formalize a few concepts we use.

²We note that in [15], time-optimality is claimed, up to factor $\tilde{\mathcal{O}}(\kappa)$. This comes as a consequence of [16], which in turn builds on [8]. However, we comment that the latter construction does not scale beyond the familiar lower bound of $\tilde{\Omega}(\sqrt{n})$, and a more precise statement would thus be that a minimum spanning tree (and thus also a GSF) requires $\tilde{\Omega}(\min\{\text{SPD}, \sqrt{n}\})$ rounds to be approximated.

2.1 The Computational Model

We follow the **CONGEST**(B) model as described in [21]. The distributed system is represented by a simple, connected weighted graph $G = (V, E, W)$, where V is the set of nodes, E is the set of edges, and $W : E \rightarrow \mathbb{N}$ is the edge weight function.³ As a convention, we use n to denote the number of nodes. We assume that all edge weights are bounded by some polynomial in n , and that each node $v \in V$ has a unique identifier of $\mathcal{O}(\log n)$ bits (we use v to denote both the node and its identifier).

Execution proceeds in global synchronous rounds, where in each round, each node take the following three steps: (1) Receive messages sent by neighbors at the previous round, (2) perform local computation, and (3) send messages to neighbors. Initially, nodes are aware only of their neighbors; input values (if any) are assumed to be fed by the environment at time 0. Output values are placed in special output-registers. In each round, each edge can carry a message of B bits for some given parameter B of the model; we assume that $B \in \Theta(\log n)$ throughout this paper.

A basic observation in this model is that we may assume, without loss of generality, that we have a broadcast facility available, as formalized in the following lemma.

Lemma 2.1 *Suppose each $v \in V$ holds $m_v \geq 0$ messages of $\mathcal{O}(\log n)$ bits each, for a total of $M \stackrel{\text{def}}{=} \sum_{v \in V} m_v$ strings. Then all nodes in the graph can receive these M messages within $\mathcal{O}(M + \text{HD})$ rounds.*

Proof: Construct a BFS tree rooted at, say, the node r with smallest identifier ($\mathcal{O}(\text{HD})$ rounds). All nodes send their messages to their parents and forward the messages received by their children to their parent as well, until the root holds all messages. Since over no edge more than M messages need to be communicated, this requires $\mathcal{O}(M + \text{HD})$ rounds. Finally all messages are broadcast over the tree, completing in another $\mathcal{O}(M + \text{HD})$ rounds. ■

In the following, we will use this lemma implicitly whenever stating that some information is “broadcast” or “announced to all nodes.”

2.2 General Concepts

We use extensively “soft” asymptotic notation that ignores polylogarithmic factors. Formally, we say that $g(n) \in \tilde{\mathcal{O}}(f(n))$ if and only if there exists a constant $c \in \mathbb{R}_0^+$ such that $f(n) \leq g(n) \log^c(f(n))$ for all but finitely many values of $n \in \mathbb{N}$. Analogously, $f(n) \in \tilde{\Omega}(g(n))$ iff $g(n) \in \tilde{\mathcal{O}}(f(n))$, $\tilde{\Theta}(f(n)) \stackrel{\text{def}}{=} \tilde{\mathcal{O}}(f(n)) \cap \tilde{\Omega}(f(n))$, $g(n) \in \tilde{o}(f(n))$ iff for each fixed $c \in \mathbb{R}_0^+$ it holds that $\lim_{n \rightarrow \infty} g(n) \log^c(f(n))/f(n) = 0$, and $g(n) \in \tilde{\omega}(f(n))$ iff $f(n) \in \tilde{o}(g(n))$.

To model probabilistic computation, we assume that each node has access to an infinite string of independent unbiased random bits. When we say that a certain event occurs “with high probability” (abbreviated “w.h.p.”), we mean that the probability of the event not occurring can be set to be less than $1/n^c$ for any desired constant c , where the probability is taken over the strings of random bits.

³We remark that our results can be easily extended to non-negative edge weights by employing appropriate symmetry breaking mechanisms.

2.3 Some Graph-Theoretic Concepts

A *path* p connecting $v, u \in V$ is a sequence of nodes $\langle v = v_0, \dots, v_k = u \rangle$ such that for all $0 \leq i < k$, (v_i, v_{i+1}) is an edge in G . Let $\text{paths}(v, u)$ denote the set of all paths connecting nodes v and u . We use the following unweighted concepts.

- The *hop-length* of a path p , denoted $\ell(p)$, is the number of edges in it.
- The *hop distance* $\text{hd} : V \times V \rightarrow \mathbb{N}_0$ is defined as $\text{hd}(v, u) := \min\{\ell(p) \mid p \in \text{paths}(v, u)\}$.
- The *hop diameter* of a graph $G = (V, E, W)$ is $\text{HD} \stackrel{\text{def}}{=} \max_{v, u \in V} \{\text{hd}(v, u)\}$.

We use the following weighted concepts.

- The *weight* of a path p , denoted $W(p)$, is its total edge weight, i.e., $W(p) \stackrel{\text{def}}{=} \sum_{i=1}^{\ell(p)} W(v_{i-1}, v_i)$.
- The *weighted distance* $\text{wd} : V \times V \rightarrow \mathbb{R}_0^+$ is defined by $\text{wd}(v, u) \stackrel{\text{def}}{=} \min\{W(p) \mid p \in \text{paths}(v, u)\}$.
- The *weighted diameter* of G is $\text{WD} \stackrel{\text{def}}{=} \max\{\text{wd}(v, u) \mid v, u \in V\}$.

The following concepts mix weighted and unweighted ones.

- Given $h \in \mathbb{N}$ and two nodes $v, u \in V$ with hop distance $\text{hd}(v, u) \leq h$, we define the *h -weighted distance* $\text{wd}_h(v, u)$ to be the weight of the lightest path connecting v and u with at most h hops, i.e., $\text{wd}_h(v, u) \stackrel{\text{def}}{=} \min\{W(p) \mid p \in \text{paths}(v, u) \text{ and } \ell(p) \leq h\}$. If $\text{hd}(v, u) > h$, we define $\text{wd}_h(v, u) \stackrel{\text{def}}{=} \infty$. (Note that wd_h does not satisfy the triangle inequality.)
- The *shortest paths diameter* of a graph, denoted SPD , is the maximal number of hops in shortest paths: $\text{SPD} \stackrel{\text{def}}{=} \max_{u, v \in V} \{\min\{\ell(p) \mid W(p) = \text{wd}(u, v)\}\}$.

Finally, given a node v and an integer $i \geq 0$, we define $\text{ball}_v(i)$ to be the set of the i nodes that are closest to v (according to wd , where identifiers are used to break symmetry): $\text{ball}_v(i) \stackrel{\text{def}}{=} \{u : |\{w : (\text{wd}(v, u), u) \leq (\text{wd}(v, w), w)\}| \leq i\}$. Note that our concept of ball differs from the usual one: we define a ball by its center and *volume*, namely the number of nodes it contains (and not by its center and radius).

We have the following immediate property.

Lemma 2.2 *Let $v, u \in V$. If $u \in \text{ball}_v(i)$ for some $i \in \mathbb{N}$ then $\text{wd}(v, u) = \text{wd}_j(v, u)$ for all $j \geq i - 1$.*

Proof: Clearly $\text{wd}(v, u) \leq \text{wd}_j(v, u) \leq \text{wd}_{i-1}(v, u)$, and it therefore suffices to show that $\text{wd}_{i-1}(v, u) = \text{wd}(v, u)$. Let $p = \langle v = v_0, v_1, \dots, v_k = u \rangle$ be a shortest path from v to u . Since edge weights are strictly positive, we have that all the k nodes v_0, \dots, v_{k-1} are strictly closer than u to v . Hence, since $u \in \text{ball}_v(i)$, we have that $i \geq k + 1$. It follows that $\text{wd}(v, u) = \text{wd}_{i-1}(v, u)$ and we are done. ■

3 Problem Statement and Lower Bounds

3.1 The Routing Problem

In the *routing table construction* problem (abbreviated *RTC*), the local input at a node is the weight of incident edges, and the output at each node v consists of (i) a unique *label* $\lambda(v)$ and (ii) a function “ next_v ” that takes a destination label λ and produces a neighbor of v , such that given the label $\lambda(u)$

of any node u , and starting from any node v , we can reach u from v by following the next pointers. Formally, the requirement is as follows. Given a start node v and a destination label $\lambda(u)$, let $v_0 = v$ and define $v_{i+1} = \text{next}_{v_i}(\lambda(u))$ for $i \geq 0$. Then for some i we must have $v_i = u$.

The performance of a solution is measured in terms of its *stretch*: A route is said to have stretch $\rho \geq 1$ if its total weight is no more than ρ times the weighted distance between its endpoints, and a solution to RTC is said to have stretch ρ if all the routes it induces have stretch at most ρ .

Variants. Routing appears in many incarnations. We list a few important variants below.

Name-independent routing. Our definition of RTC allows for node relabeling. This is the case, as mentioned above, in the Internet. The case where no such relabeling is allowed (which can be formalized by requiring λ to be the identity function), is called *name-independent routing*.

It can be shown that assigning new labels to the nodes is unavoidable by proving that any (randomized) algorithm achieving polylogarithmic (expected) stretch without relabeling must run for $\tilde{\Omega}(n)$ rounds. Formally, we can prove the following.

Theorem 3.1 *In the **CONGEST** model, any algorithm for RTC that produces name-independent stateful routing with expected average stretch ρ requires $\Omega(n/(\rho^2 \log n))$ time.*

Stateful routing. The routing problem as defined above is *stateless* in the sense that routing a packet is done regardless of the path it traversed so far. One may also consider *stateful* routing, where while being routed, a packet may gather information that helps it navigate later (one embodiment of this idea in the Internet routing today is MPLS, where packets are temporarily piggybacked with extra headers). Note that the set of routes to a single destination in stateless routing must constitute a tree, whereas in stateful routing even a single route may contain a cycle. Formally, in stateful routing the label of the destination may change from one node to another: The next_v function outputs both the next hop (a neighbor node), and a new label λ_v used in the next hop.

Name-independent routing. Our definition of RTC allows for node relabeling. This is the case, as mentioned above, in the Internet. The case where no such relabeling is allowed (which can be formalized by requiring λ to be the identity function), is called *name-independent routing*.

It can be shown that assigning new labels to the nodes is unavoidable by proving that any (randomized) algorithm achieving polylogarithmic (expected) stretch without relabeling must run for $\tilde{\Omega}(n)$ rounds. What might come as a surprise here is that the result also applies to stateful routing.

Theorem 3.2 *In the **CONGEST** model, any algorithm for RTC that produces name-independent routing with (expected) average stretch ρ requires $\Omega(n/(\rho^2 \log n))$ time.*

3.2 The Distance Approximation Problem

The *distance approximation* problem is akin to the routing problem. Again, each node v outputs a label $\lambda(v)$, but now, v needs to construct a function $\text{dist}_v : \lambda(V) \rightarrow \mathbb{R}^+$ (the table) such that for all $w \in V$ it holds that $\text{dist}_v(w) \geq \text{wd}(v, w)$. The stretch of the approximation for a given node w is $\text{dist}_v(w)/\text{wd}(v, w)$, and the solution has stretch $\rho \geq 1$, if $\text{dist}_v(w) \leq \rho \text{wd}(v, w)$ for all $v, w \in V$.

Similarly to routing, we call a scheme name-independent if λ is the identity function. Since we

require distances estimates to be produced without communication, there is no “stateful” distance approximation.

3.3 Hardness of Name-Independent Distributed Table Construction

While name-independence may be desirable, our routing and distance approximation algorithm makes heavy use of relabeling. This is unavoidable for fast construction, because, as the following two theorems show, any name-independent scheme of polylogarithmic stretch requires $\tilde{\Omega}(n)$ rounds for table construction. The lower bound holds even for stateful routing and average stretch. Moreover, since the construction below is generic, intuitively it implies that there is no reasonable restriction, be it in terms of topology, edge weights, or node degrees, that permits fast construction of name-independent routing tables.⁴

Theorem 3.3 *In the **CONGEST** model, any name-independent routing scheme of (expected) average stretch ρ requires $\Omega(n/(\rho^2 \log n))$ rounds for table construction. This holds even if all edge weights are 1, the graph is a tree of constant depth, and the node identifiers are $1, \dots, n$.*

Proof: We assume w.l.o.g. that all set sizes we use in this proof are integer and that nodes may send no more than exactly $\log n \in \mathbb{N}$ bits over each edge in each round. Consider the following family of trees of depth 2. The root is connected to $n_1 \in \Theta(\rho)$ inner nodes, each of which has n_2 children; denote by I and L the respective sets of nodes. All edges have weight 1, i.e., the maximal simple path weight is 4.

We assign the identifiers $1, \dots, n$ uniformly at random to the $n_1 n_2$ leaves (w.l.o.g., we neglect that the total number of nodes is $n_1 n_2 + n_1 + 1$ in the following and use n instead). Consider any deterministic algorithm constructing routing tables within $r \in \mathbb{N}$ rounds. From each node in I , the root receives at most $r \log n$ bits, hence there are at most $2^{r n_1 \log n}$ possible routing tables at the root. Now consider the $n!/(n_2!)^{n_1}$ possible partitions of the leaf identifiers to the subtrees rooted at nodes from I . We bound the number of such partitions for which a fixed routing table at the root may serve a uniformly random routing request with probability at least p correctly. This requirement translates to at least pn identifiers being exactly in the subtree where the routing table points to; we have $\binom{n}{pn}$ possible choices for these identifiers. The remaining $(1-p)n$ identifiers may be distributed arbitrarily to the remaining subtrees. Depending on the distribution of the pn identifiers we already selected, the number of possibilities for this may vary. Using standard arguments it can be shown that this quantity is maximized if the pn identifiers are distributed evenly among the subtrees, i.e., each of them contains pn_2 of them. We conclude that no routing table can serve a uniform request with probability at least p for more than $\binom{n}{pn}((1-p)n)!/((1-p)n_2!)^{n_1}$ of the possible input partitions. Considering the number possible routing tables and the total number of input partitions $n!/(n_2!)^{n_1}$, we have that

$$\frac{(pn)!((1-p)n_2!)^{n_1}}{n_2!^{n_1}} = \frac{n!/n_2!^{n_1}}{\binom{n}{pn}((1-p)n)!/((1-p)n_2!)^{n_1}} \leq 2^{r n_1 \log n}.$$

⁴The lower bound graph can be adapted to be a balanced binary tree, weakening the lower bound on the stretch by factor $\log n$.

We distinguish two cases, the first being $p < e^2/n_1$. We seek to upper bound p in the second case as well, where $p \geq e^2/n_1$. Clearly the l.h.s. of the above inequality is increasing in $p \in [0, 1]$. Together with Stirling's approximation $x! \in e^{(1-o(1))x(\ln x - 1)}$ we can bound

$$\begin{aligned}
\frac{(pn)!((1-p)n_2!)^{n_1}}{n_2!^{n_1}} &\geq \frac{(e^2 n_2)!((1-e^2/n_1)n_2)^{n_1}}{n_2!^{n_1}} \\
&\in e^{(1-o(1))(e^2 n_2(\ln(e^2 n_2)-1)+n(1-e^2/n_1)(\ln((1-e^2/n_1)n_2)-1)-n(\ln n_2-1))} \\
&\subseteq e^{(1-o(1))(e^2 n_2(\ln n_2+1)-e^2 n_2(\ln n_2-1)+n \ln(1-e^2/n_1))} \\
&\subseteq e^{(1-o(1))(2e^2 n_2-e^2 n_2)} \\
&= e^{(e^2-o(1))n_2}.
\end{aligned}$$

The assumption that $p \geq e^2/n_1$ thus implies (for sufficiently large n) that $r > n_2/(n_1 \log n)$.

Now condition on the event that for the given routing request the table does not lead to the correct subtree. We fix the (uniformly random) subset of leaf identifiers in the subtree S the root's routing table points to, and conclude that the set of remaining identifiers is a uniformly random subset of $n - n_2 - 1$ leaf identifiers plus the destination's identifier. Moreover, the destination is uniformly random from this subset and the remaining identifiers are uniformly distributed among the remaining subtrees. We delete S from the graph (since clearly there is no reason to route to S again) and examine the next routing decision of the root. We observe that the situation is identical to the initial setting except that n_1 is replaced by $n_1 - 1$. Note also that S contained no valuable information: We deleted S and the identifiers in S from the graph, and any other information known to nodes in S must have been communicated to S by the root. Hence, repeating the above arguments, we see that the probability to find the destination in the second attempt conditioned on the first having failed is at most $e^2/(n_1 - 1)$ or $r > n_2/(n_1 \log n)$. By induction on the number of routing attempts, we infer that for $i \in \{1, \dots, n_1/2\}$, the probability p_i to succeed in the i^{th} attempt to route from the root node to the subtree containing the destination (conditional on the previous attempts having failed) is upper bounded by $2e^2/n_1$ unless $r > n_2/(n_1 \log n)$.

Overall, the probability that a deterministic algorithm constructing routing tables within $r \leq n_2/(n_1 \log n)$ rounds fails to serve a uniformly random routing request at the root for uniformly distributed leaf identifiers using fewer than $n_1/2$ attempts (i.e., visits of the root on the routing path) is lower bounded by

$$\left(1 - \frac{2e^2}{n_1}\right)^{n_1/2} \in \Omega(1).$$

Note that an analogous argument holds for routing requests issued at other nodes, since they have a large probability to require routing to a different subtree. Therefore, the average stretch of any deterministic routing algorithm running for fewer than $n_2/(n_1 \log n)$ rounds is at least $\Omega(n_1)$. By Yao's principle, the expected average stretch of randomized algorithms running for fewer than $n_2/(n_1 \log n)$ rounds thus must also be in $\Omega(n_1)$. Recalling that $n_1 \in \Theta(\rho)$ and $n_2 = n/\rho$, we get that $r \in o(n/(\rho^2 \log n))$ rounds are insufficient to achieve (expected) average stretch ρ , proving the statement of the theorem. ■

A streamlined version of the argument shows that a similar lower bound applies to distance approximation.

Theorem 3.4 *In the **CONGEST** model, any name-independent distance approximation scheme of (expected) average stretch ρ requires $\Omega(n/\log n)$ rounds for table construction in graphs with edge weights of 1 and $\omega_{\max} \in \mathcal{O}(\rho)$ only. This holds even if the graph is a star and the node identifiers are $1, \dots, n$.*

Proof: Again, we assume w.l.o.g. that all considered values are integer and that link capacity is $\log n$ bits per round. Suppose G is a star with n leafs (we neglect w.l.o.g. the center in the node count). All edges have weight ω_{\max} with independent probability $1/2$; the remaining edges have weight 1.

Condition on the event that some fixed node's v incident edge has weight 1. Thus, there are two possible path weights to other nodes: $\omega_{\max} + 1$ and 2. Within r rounds, the node receives at most $r \log n$ bits, yielding $2^{r \log n}$ possible distance estimate configurations. In order to be ρ -approximate for $\rho < (\omega_{\max} + 1)/2$ and a given other leaf, v 's table must output an estimate of at most $2\rho < \omega_{\max} + 1$ in case the leaf's edge has weight 2 and at least $\omega_{\max} + 1$ if the leaf's edge has weight ω_{\max} . Thus any given table can be correct for a given leaf for only one of the two possible choices of the leaf's edge's weight. There are 2^n possible edge weight assignments. By the above observation, a fixed table is ρ -approximate for a given destination with probability $1/2$. By Chernoff's bound, this implies that the probability that a fixed table is correct for a fraction of $3/4$ of the destinations is bounded by $2^{-\Omega(n)}$. By the union bound, it follows that for the given uniformly random edge weight assignment, the probability that the computed table is correct for a fraction of $3/4$ of the destinations is upper bounded by $2^{-\Omega(n)} 2^{r \log n}$. This implies that $r \in \Omega(n/\log n)$ or the average stretch of node v 's table must be $\Omega(\omega_{\max})$.

By symmetry, the same applies to all nodes incident to an edge of weight 1. By Chernoff's bound, w.h.p. at least one quarter of the nodes satisfies this property, i.e., the probability mass of the events where fewer than $n/4$ edges have weight 1 is negligible. By linearity of expectation, it follows that any deterministic algorithm running for $o(n/\log n)$ rounds exhibits average stretch $\Omega(\omega_{\max})$, and by Yao's principle this extends to the expected stretch of randomized algorithms. ■

Consequently, in the remainder of the paper we shall consider name-dependent schemes only.

3.4 Hardness of Diameter Estimation

In [12], it is shown that approximating the hop-diameter of a network within a factor smaller than 1.5 cannot be done in the **CONGEST** model in $\tilde{o}(\sqrt{n})$ time. Here, we prove a hardness result for the weighted diameter, formally stated as follows.

Theorem 3.5 *For any $\omega_{\max} \geq \sqrt{n}$, there is a function $\alpha(n) \in \Omega(\omega_{\max}/\sqrt{n})$ such that the following holds. In the family of weighted graphs of hop-diameter $\text{HD} \in \mathcal{O}(\log n)$ and edge weights 1 and ω_{\max} only, an (expected) $\alpha(n)$ -approximation of the weighted diameter requires $\tilde{\Omega}(\sqrt{n})$ communication rounds in the **CONGEST** model.*

Proof sketch: We construct a graph G_n with $\Theta(n)$ nodes. Let $m = \sqrt{n} \in \mathbb{N}$. The graph consists of the following three conceptual parts. Figure 1 illustrates a part of the construction.

- Nodes $v_{i,j}$ for $1 \leq i, j \leq m$. These nodes are connected as m paths of length $m - 1$. All path edges are of weight 1.

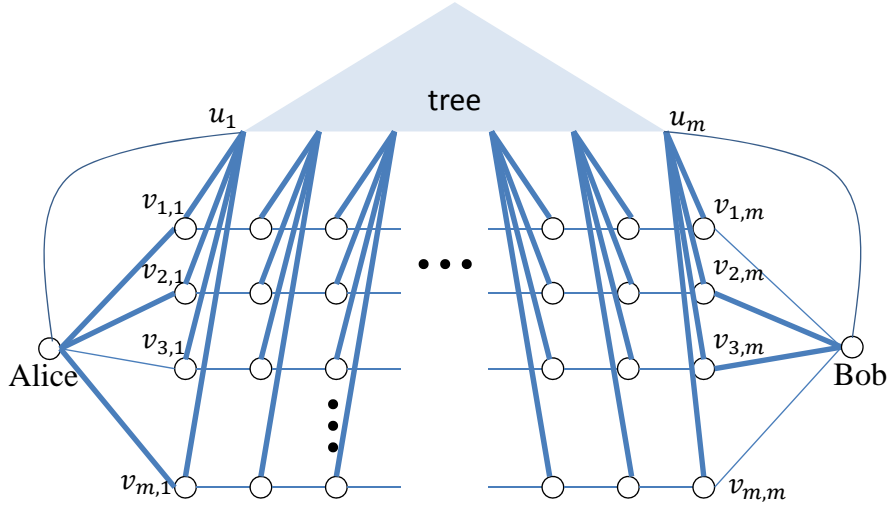


Figure 1: An illustration of the graph used in the proof of [Theorem 3.5](#). Thick edges denote edges of weight ω_{\max} , other edges are of weight 1. The shaded triangle represents a binary tree

- A star rooted at an *Alice* node, where the children are $v_{1,1}, \dots, v_{m,1}$, and similarly, a star rooted at a *Bob* node, whose leaves are $v_{m,1}, \dots, v_{m,m}$. We specify the weights of these edges later.
- For each $1 \leq j \leq m$ there is a node u_j connected to all nodes $v_{i,j}$, $1 \leq i \leq m$ in “column” j , with edges of weight ω_{\max} . In addition, there is a binary tree whose leaves are the nodes u_j . All tree edges have weight 1. Finally, *Alice* and *Bob* are connected to u_1 and u_m , respectively, by edges of weight 1.

It is easy to see that the hop-diameter of G_n is $\mathcal{O}(\log n)$: the hop-distance from any node to one of the nodes u_j is $\mathcal{O}(\log n)$, and the distance between any two such nodes is also $\mathcal{O}(\log n)$. However, the majority of the short paths guaranteeing the small diameter passes through very few nodes close to the root of the binary tree. Consequently, it takes a long time to exchange a large number of bits between *Alice* and *Bob*, implying that it is hard to decide set disjointness for sets held by *Alice* and *Bob* in the **CONGEST** model. Specifically, the following fact is a direct corollary from [7].

Fact 3.1 ([7]) *Let $\mathcal{M} \stackrel{\text{def}}{=} \{1, \dots, m\}$. Suppose that node *Alice* holds a set $A \subseteq \mathcal{M}$ and that node *Bob* holds a set $B \subseteq \mathcal{M}$. Then finding whether $A \cap B = \emptyset$ takes $\tilde{\Omega}(m)$ rounds in the **CONGEST** model, even for randomized algorithms.*

We now show that if the diameter of G_n can be approximated within factor ω_{\max}/\sqrt{n} in time T in the **CONGEST** model, then the set disjointness problem can be solved in time $T + 1$. To this end, we set the edge weights of the stars rooted at *Alice* and *Bob* as follows: for all $i \in \{1, \dots, m\}$, the edge from *Alice* to $v_{i,1}$ has weight ω_{\max} if $i \in A$ and weight 1 else; likewise, the edge from *Bob* to $v_{i,m}$ has weight ω_{\max} if $i \in B$ and weight 1 else.

Note that given A at *Alice* and B at *Bob*, we can inform the nodes $v_{i,1}$ and $v_{i,m}$ of these weights in one round. Now run any algorithm that outputs a value between WD and $\alpha(n)\text{WD} \stackrel{\text{def}}{=} \omega_{\max}\text{WD}/(\sqrt{n} + C \log n)$ (for a suitable constant C) within T rounds, and output “ A and B are disjoint” if the outcome is at most ω_{\max} and output “ A and B are not disjoint” otherwise.

It remains to show that the outcome of this computation is correct for any inputs A and B and the statement of the theorem will follow from [Fact 3.1](#) (recall that the number of nodes of G_n is $\Theta(n)$). Suppose first that $A \cup B = \emptyset$. Then for each node $v_{i,j}$, there is a path of at most \sqrt{n} edges of weight 1 connecting it to *Alice* or *Bob*, and *Alice* and *Bob* are connected to all nodes in the binary tree and each other via $\mathcal{O}(\log n)$ hops in the binary tree (whose edges have weight 1 as well). Hence the weighted diameter of G_n is $\sqrt{n} + \mathcal{O}(\log n)$ in this case and the output is correct (where we assume that C is sufficiently large to account for the $\mathcal{O}(\log n)$ term). Now suppose that $i \in A \cap B$. In this case each path from node $v_{i,1}$ to *Bob* contains an edge of weight ω_{\max} , since the edges from *Alice* to $v_{i,1}$ and *Bob* to $v_{i,m}$ as well as those connecting $v_{i,j}$ to u_j have weight ω_{\max} . Hence, the weighted distance from $v_{i,1}$ to *Bob* is strictly larger than ω_{\max} and the output is correct as well. This shows that set disjointness is decided correctly and therefore the proof is complete. ■

3.5 Hardness of Name-Dependent Distributed Table Construction

A lower bound on name-dependent distance approximation follows directly from [Theorem 3.5](#).

Corollary 3.6 *For any $\omega_{\max} \geq \sqrt{n}$, there is a function $\alpha(n) \in \Omega(\omega_{\max}/\sqrt{n})$ such that the following holds. In the family of weighted graphs of hop-diameter $\text{HD} \in \mathcal{O}(\log n)$ and edge weights 1 and ω_{\max} only, constructing labels of size $\tilde{o}(\sqrt{n})$ and tables for distance approximation of (expected) stretch $\alpha(n)$ requires $\tilde{\Omega}(\sqrt{n})$ communication rounds in the **CONGEST** model.*

Proof: We use the same construction as in the previous proof, however, now we need to solve the disjointness problem using the tables and labels. Using the same setup, we run the assumed table and label construction algorithm. Afterwards, we transmit, e.g., the label of *Alice* to all nodes $v_{i,1}$. This takes $\tilde{o}(\sqrt{n})$ rounds due to the size restriction of the labels. Then we query the estimated distance to *Alice* at the nodes $v_{i,1}$ and collect the results at *Alice*. Analogously to the proof of [Theorem 3.5](#), the maximum of these values is large if and only if the input satisfies that $A \cap B = \emptyset$. Since transmitting the label costs only $\tilde{o}(\sqrt{n})$ additional rounds, the same asymptotic lower bound as in [Theorem 3.5](#) follows. ■

A variation of the theme shows that stateless routing requires $\tilde{\Omega}(\sqrt{n})$ time.

Corollary 3.7 *For any $\omega_{\max} \geq \sqrt{n}$, there is a function $\alpha(n) \in \Omega(\sqrt{\omega_{\max}}/n)$ such that the following holds. In the family of weighted graphs of hop-diameter $\text{HD} \in \mathcal{O}(\log n)$ and edge weights 1 and ω_{\max} only, constructing stateless routing tables of (expected) stretch $\alpha(n)$ with labels of size $\tilde{o}(\sqrt{n})$ requires $\tilde{\Omega}(\sqrt{n})$ communication rounds in the **CONGEST** model.*

Proof sketch: We consider the same graph as in the proof of [Theorem 3.5](#) and input sets A and B at *Alice* and *Bob*, respectively, but we use a different assignment of edge weights.

- All edges incident to a node in the binary tree have weight ω_{\max} .
- For each $i \in \{1, \dots, m\}$, the edge from *Alice* to $v_{i,1}$ has weight ω_{\max} if $i \in A$ and weight 1 else. Likewise, the edge from *Bob* to $v_{i,m}$ has weight ω_{\max} if $i \in B$ and otherwise weight 1.
- The remaining edges (on the m paths from $v_{i,1}$ to $v_{i,m}$) have weight 1.

Observe that the distance from *Alice* to *Bob* is $\sqrt{n} + 1$ if $A \cap B \neq \emptyset$ and strictly larger than ω_{\max} if $A \cap B = \emptyset$. Once static routing tables for routing on paths of stretch at most $\omega_{\max}/(\sqrt{n} + 1)$ are set

up, e.g. *Bob* can decide whether A and B are disjoint as follows. *Bob* sends its label to *Alice* via the binary tree (which takes time $\tilde{o}(\sqrt{n})$ if the label has size $\tilde{o}(\sqrt{n})$). *Alice* responds with “ i ” if the first routing hop from *Alice* to *Bob* is node $v_{i,1}$ and $i \in A$ (i.e., the weight of the edge is 1), and “ $A \cap B = \emptyset$ ” else (this takes $\mathcal{O}(\log n)$ rounds). *Bob* then outputs “ $A \cap B \neq \emptyset$ ” if *Alice* responded with “ i ” and $i \in B$ (i.e., the weight of the routing path is $\sqrt{n} + 1$ since the edge from *Bob* to $v_{i,m}$ has weight 1) and “ $A \cap B = \emptyset$ ” otherwise.

If the output is “ $A \cap B \neq \emptyset$ ”, it is correct because $i \in A \cap B$. On the other hand, if it is “ $A \cap B = \emptyset$ ”, the route from *Alice* to *Bob* must contain an edge of weight ω_{\max} , implying by the stretch guarantee that there is no path of weight $\sqrt{n} + 1$ from *Alice* to *Bob*. This in turn entails that $A \cap B = \emptyset$ due to the assignment of weights and we conclude that the output is correct also in this case. Hence the statement of the corollary follows from [Fact 3.1](#). ■

We remark that [Theorem 3.4](#), [Theorem 3.5](#), [Corollary 3.6](#), and [Corollary 3.7](#) have in common that if edge weight 0 is permitted, no stretch bound faster than the stated lower bounds even if the only other feasible edge weight is 1.

Finally, we note that the hop-diameter is also an obvious lower bound on the time required to approximate the weighted diameter, construct stateless routing tables, etc. since if the running time is smaller than HD, distant parts of the graph (in the sense of hop-distance) cannot influence the local output.

4 Routing Algorithm

Overview. To construct routing tables, one needs to learn about paths. Naïve distributed algorithms explore paths sequentially, adding one edge at a time, leading to potentially linear complexity, since shortest weighted paths may be very long in terms of the number of edges. Our basic idea is to break hop-wise long paths into small pieces by means of random sampling. Specifically, motivated by the $\tilde{\Omega}(\sqrt{n})$ lower bound of [Theorem 3.5](#), we select a random subset of $\tilde{\Theta}(\sqrt{n})$ nodes we call the *skeleton*. It follows that, w.h.p., (1) any simple path of hop-length $\tilde{\Omega}(\sqrt{n})$ contains a skeleton node, and (2) any node has a skeleton node among its closest $\tilde{\mathcal{O}}(\sqrt{n})$ nodes. The route that our scheme will select from a given source to a given destination depends on their distance: If the destination is one of the $\tilde{\mathcal{O}}(\sqrt{n})$ nodes closest to the destination, routing will be done using a “short range scheme” (see below); otherwise, the short range scheme is used to route from the source to the nearest skeleton node, from which, using another scheme we call “long distance routing,” we route to the skeleton node closest to the destination node, and finally, another application of the short range scheme brings us to the destination. Intuitively, we can split the problem into the following tasks:

1. Short range scheme: how to route efficiently from each node to its $\tilde{\Theta}(\sqrt{n})$ closest nodes including at least one skeleton node, and, conversely, from a skeleton node to all its “subordinates” (note the asymmetry in this case).
2. Skeleton routing scheme: how to route between skeleton nodes efficiently.

The short range scheme is described in [Section 4.2](#). We note that since a straightforward application of multiple-source shortest paths may result in linear time, we develop a hierarchical structure to

Algorithm 1: $\text{BSP}(h, \Delta, S)$: Bounded shortest paths, computed at node $v \in V$.

```

input      :  $h$                                 //range parameter: hop bound on path lengths, globally known
               $\Delta$  //overlap parameter: number of closest sources each node needs to detect, globally known
               $\text{source} : V \rightarrow \text{SID} \cup \{\perp\}$  //each  $v$  knows  $\text{source}(v)$ ;  $\text{source}(v) = \perp$  means  $v$  is not a source
computes: For all  $t \in \{1, \dots, h\}$ :  $t$ -weighted distance and the next hop from  $v$  to each of the closest  $\Delta$ 
              source node sets using paths of at most  $t$  edges (or all such sets, if there are at most  $\Delta$  within
               $t$  hops).
1 if  $\text{source}(v) \neq \perp$  then  $L_v(0) := \{(0, \text{source}(v), v)\}$  else  $L_v(0) := \emptyset$                                 //initialization
2 for  $t := 1$  to  $h$  do
3   send  $L_v(t-1)$  to all neighbors;  $L_v(t) := \emptyset$ 
4   foreach neighbor  $u$  do
5     receive  $L_u(t-1)$ 
6     foreach  $(d_u, s_u, \text{next}_u) \in L_u(t-1)$  do                                //Bellman-Ford relaxation
7       if  $\exists (d_v, s_v, \text{next}_v) \in L_v(t)$  s.t.  $s_v = s_u$  then
8         if  $(d_u + W(u, v), u) < (d_v, \text{next}_v)$  then                                //comparisons are lexicographical
9            $L_v(t) := L_v(t) \setminus \{(d_v, s_v, \text{next}_u)\} \cup \{(d_u + W(u, v), s_u, u)\}$ 
10        else  $L_v(t) := L_v(t) \cup \{(d_u + W(u, v), s_u, u)\}$ 
11   truncate  $L_v(t)$  to smallest  $\Delta$  entries                                //order is lexicographical
12 return  $(L_v(1), \dots, L_v(t))$ 

```

solve the short-range routing. This hierarchy bears resemblance to the Thorup-Zwick distance oracle algorithm [30]. Our long distance routing is described in Section 4.3. The main challenge there is to build the skeleton graph; since it might be too dense, we sparsify it “on the fly” while constructing it. This construction is implemented by adapting the spanner algorithm of Baswana and Sen [3] to our setting.

We start by describing the variant of the Bellman-Ford algorithm we use as a basic building block in Section 4.1.

4.1 Bounded Shortest Paths

We now describe a basic subroutine we use. Algorithm BSP, whose pseudo code is given in Algorithm 1, is essentially a standard multiple-source distributed Bellman-Ford algorithm, with two restrictions: first, the algorithm is run for only h rounds (cf. Line 2); and second, nodes never report more than Δ sources (cf. Line 11).

We consider a slightly extended variant of the algorithm: In the original algorithm, each node is a “source” and the goal is to compute the distances of all nodes to it. Here we assume that (i) not all nodes are sources, and (ii) sets of nodes may act as a single source, as if there were 0-weight edges connecting them. Both extensions are modeled by the source function, that maps a node to \perp if it is not a source, or multiple nodes to the same source ID if they are in the same source set. We use S to denote the set of *sources*, i.e., $S = \{\text{source}(v) \mid v \in V\} \setminus \{\perp\}$, and for each $s \in S$, the *source nodes* of s is $SN(s) := \{v \mid \text{source}(v) = s\}$. Note that the source function uniquely determines the source sets and vice versa. We assume that a source ID can be encoded using $\mathcal{O}(\log n)$ bits.

We analyze the algorithm leveraging the correctness of the basic Bellman-Ford algorithm. To this end, let us define [Algorithm 1*](#) by omitting [Line 11](#) from [Algorithm 1](#) and fixing $h = n - 1$. Observing that [Algorithm 1*](#) is exactly the distributed Bellman-Ford algorithm, we may conclude the following standard property.

Lemma 4.1 *Fix an execution of [Algorithm 1*](#). Denote by $L_v^*(t)$ for some $0 \leq t \leq n - 1$ and $v \in V$ the contents of the L_v variable at node v after t iterations of [Algorithm 1*](#). Then for each (d, s, next) entry in $L_v^*(t)$ we have that s is a source and $\text{wd}_t(v, s) := \min_{u \in SN(s)} \{\text{wd}_t(v, u)\} = d$, namely d is the length of the shortest path that consists of at most t edges from v to any node u in the source set of s . Moreover, next is the next node on that shortest path from v to u .*

[Lemma 4.1](#) says that running only h iterations is sufficient if we are interested in paths of h or less edges only. We now consider the effect of repeatedly truncating the distance vector.

Lemma 4.2 *Consider executions of [Algorithm 1](#) and of [Algorithm 1*](#) on the same graph and with the same source function. Let $L_v(t)$ and $L_v^*(t)$ denote the contents of the L_v variable at node v after t iterations under [Algorithm 1](#) and under [Algorithm 1*](#), respectively. Then $L_v(t)$ contains exactly the smallest Δ entries of $L_v^*(t)$ with respect to lexicographical ordering (or the entire list, if $|L_v^*(t)| \leq \Delta$).*

Proof: By induction on t . The base case is $t = 0$, and the lemma clearly holds upon initialization ([Line 1](#)). For the induction step, assume that the lemma holds for $t - 1 \in \{0, \dots, h - 1\}$ at all nodes, and consider iteration t at some node v . By the induction hypothesis, we have that the message received by v from each neighbor u at time t under [Algorithm 1](#) is exactly the top Δ entries sent by node u at time t under [Algorithm 1*](#), because these entries are computed at the end of iteration $t - 1$. The lemma therefore follows from the fact that for any $k \geq 0$, the smallest k entries of a union of sets are contained in the union of the smallest k entries from each set. ■

Note that the information provided by $L_v(h)$ is insufficient for routing: since the Δ closest source node sets may differ between neighbors, it may be the case that for some source identifier s and two neighbors v and u we have that u is the next node from v to s in $L_v(h)$, but there is no entry for source s in $L_u(h)$! This occurs, for example, if in iteration h , u learns about a source set closer than $SN(s)$, pushing s out of $L_u(h)$. However, since the algorithm returns $(L_v(1), \dots, L_v(h))$ instead of simply $L_v(h)$, we can still reconstruct the detected paths.

Lemma 4.3 *For any node v and any entry $(d, s, \text{next}) \in L_v(h)$, a routing path of at most h hops from v to a node in s of weight d can be constructed using the L tables at the nodes and a hop counter.*

Proof: The routing decision for hop t at the current node v_{t-1} (where $v_0 := v$) is made by looking up the entry $(d_{t-1}, s, \text{next}) \in L_{v_{t-1}}(h - (t - 1))$. We show by induction on the length $\ell \leq h$ of a shortest path from v to its closest node $u \in SN(s)$ that such an entry always exists. Note that by [Lemmas 4.1](#) and [4.2](#), such an entry satisfies that $d_{t-1} = \text{wd}_{h-(t-1)}(v_{t-1}, u)$ and thus the constructed path has weight $\text{wd}_h(v, u) = d$. Trivially, the claim is true for $\ell = 0$ by initialization of the lists $L_u(0)$, $u \in V$.

Now suppose the claim holds for $\ell \in \mathbb{N}_0$ and consider node v with entry $(\text{wd}(v, u), s, \text{next}) \in L_v(h)$. Suppose w is the neighbor of v which is next on the shortest ℓ -hop path from v to u . Hence it is the endpoint of a shortest $(\ell - 1)$ -hop path from w to u , and there is no shorter path from w to any node in $SN(s)$ of at most $\ell - 1$ hops (otherwise there would be a shorter path of at most ℓ hops from

v to a node in $SN(s)$). Therefore, by Lemma 4.1, $(\text{wd}_{h-1}(w, u), s, \text{next}_w) \in L_u^*(h-1)$ for some next_w . Assuming for contradiction that $(\text{wd}_{h-1}(w, u), s, \text{next}_w) \notin L_u(h-1)$ implied, by Lemma 4.2, that there are Δ entries $(d, s', \text{next}) \in L_u(h-1)$ that are lexicographically smaller than $(\text{wd}_{h-1}(w, u), s, \text{next}_w)$. Node w would send these smaller entries in iteration h of Algorithm 1, yielding the contradiction that $(\text{wd}(v, u), s, \text{next}_w) \notin L_v(h)$. It follows that indeed $(\text{wd}_{h-1}(w, u), s, \text{next}_w) \in L_u(h-1)$ and the proof concludes. ■

We summarize the properties of Algorithm 1 with the following theorem.

Theorem 4.4 *Algorithm 1 computes the h -weighted distance and next hop of a shortest path of at most h edges from each node to its closest Δ source sets. Each node on the corresponding shortest path can determine the next hop on the path out of the number of preceding hops and the output of the algorithm. The time complexity of Algorithm 1 in the CONGEST model is $\mathcal{O}(\Delta h)$ rounds.*

Proof: Correctness follows from Lemmas 4.1 and 4.2. Lemma 4.3 proves that the paths can be reconstructed as stated. The time complexity follows from the fact that the algorithm runs for h iterations, and each iteration can be implemented in $\mathcal{O}(\Delta)$ rounds in the CONGEST model since the messages contain $\mathcal{O}(\Delta)$ IDs and distances. ■

Stateless routing. The routing mechanism suggested by Lemma 4.3 has the disadvantage that it is stateful, as the routing decision depends on the number of previous routing hops. It is easy to make it stateless: at each node, a packet is directed toward the hop that reported the best distance estimate, i.e., the next hop to take at node v for destination s is $\arg \min_{\text{next}} \{d : (d, s, \text{next}) \in \bigcup_t L_v(t)\}$.

Corollary 4.5 *For any node v and any entry $(d, s, \text{next}) \in L_v(h)$, a routing path of at from v to a node in s of weight d can be constructed using the local knowledge of the nodes only.*

Proof: Lemma 4.3 shows that if a node w follows the next_w pointer of any entry $(d_w, s, \text{next}_w) \in L_w(t)$ for any $t \in \{1, \dots, h\}$, node next_w has an entry $(d' - W(w, \text{next}_w), s, \text{next}_{\text{next}_w}) \in L_w(t-1)$. We thus can simply choose to follow at each node v the next pointer of entry $(d, s, \text{next}) \in \bigcup_{t \in \{1, \dots, h\}} L_v(t)$ with minimal d and are guaranteed to eventually arrive at some node in s using a path of weight at most d . ■

Note that in general we cannot guarantee that the constructed path has at most h hops when applying this mechanism; this holds true, however, if we are routing to one of the h nodes closest to the source of the routing request (by Lemma 2.2). This observation will be crucial for making our general routing scheme stateless.

4.2 The Short-Range Scheme

With Algorithm BSP at hand, we can now describe our short-range routing scheme. Our goal is to allow each node to find a route to each of its closest $\tilde{\Theta}(\sqrt{n})$ neighbors. A naïve application of Algorithm BSP, where all nodes are sources, would set the overlap parameter to $\tilde{\Theta}(\sqrt{n})$ (this is the number of nodes we want to know about), and the range parameter to $\tilde{\Theta}(\sqrt{n})$ too (in order to find the closest $\tilde{\Theta}(\sqrt{n})$ nodes it suffices to go to this hop-distance, cf. Lemma 2.2). However, Theorem 4.4 tells us that in this case, the time complexity would be $\mathcal{O}(\Delta h) \subset \tilde{\mathcal{O}}(n)$, a far cry from the $\tilde{\Omega}(\sqrt{n})$ lower bound from Corollaries 3.6 and 3.7. Our solution is a hierarchical bootstrapping process that

converges in double-exponential speed. We show that the stretch is proportional to the number of stages in the hierarchy.

The Construction

The construction is done iteratively in L stages. In the interest of clarity we describe the construction intuitively first and then formalize it. The idea is that on the one hand we want to spend at most a certain amount of time, but on the other hand with each stage try to reduce the number of landmarks as quickly as possible. This approach is the spirit of Thorup-Zwick distance oracles and routing schemes [29, 30], and it is also used in a distributed fashion in [6]. The difficulty lies in constructing such a hierarchy quickly.⁵

The sets of landmarks, denoted S_1, \dots, S_L , are sampled uniformly and independently at random without any coordination overhead, with $S_0 \stackrel{\text{def}}{=} V$, and $S_i \subseteq S_{i-1}$ for $1 \leq i \leq L$. In the i^{th} stage, each node finds a route to the closest node in S_i as well as to all nodes in S_{i-1} that are closer to it. This property allows us to bound the routing stretch. The basic argument is a simple application of the triangle inequality (see Figure 2): Consider a route from node v to node w . If there is a node $u \in S_1$ that is closer to v than w , then the route of shortest paths via u has stretch at most 3. It is therefore sufficient for v to determine (the next hop of) least-weight routes to nodes in $S_0 \stackrel{\text{def}}{=} V$ that are closer to it than the closest node in S_1 only. Using double induction, we can bound the stretch of the multi-stage application of this technique we employ.

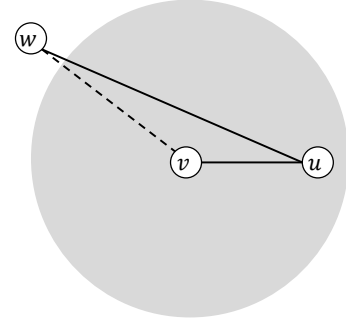


Figure 2: The distance from v to w is at least one third of the length of the route from v to w via u .

To this end, in each stage we invoke Algorithm 1 with source set S_{i-1} . We now explain how to choose the parameters h_i and Δ_i for this invocation. Let p_i be the probability of a node to be selected into S_i . Then w.h.p., each node v has a member of S_i among the $\mathcal{O}(\log n/p_i)$ nodes closest to v . Hence, this is a good choice for the distance parameter h_i . The expected number of nodes from S_{i-1} among the h_i nodes closest to a given node is $p_{i-1}h_i$. Applying Chernoff's bound shows that this number is bounded by $\mathcal{O}(p_{i-1}h_i) = \mathcal{O}(p_{i-1} \log n/p_i)$ w.h.p. This is an upper bound on the number of sources that need to be detected by each node and therefore is our choice of the overlap parameter Δ_i .

The resulting running time of the call to Algorithm BSP is $\mathcal{O}(h_i \Delta_i) \subset \tilde{\mathcal{O}}(p_{i-1}/p_i^2)$. Since this is the dominating term in the running time in each stage, it is now easy to determine the sampling probabilities: neglecting polylogarithmic factors, we get the simple recursion $p_i = \sqrt{p_{i-1}/T}$, where T is the desired running time and $p_0 \stackrel{\text{def}}{=} 1$. For example, if we want to ensure a running time bound of $T \in \tilde{\mathcal{O}}(\sqrt{n})$, we obtain:

⁵In [6], distance sketches are constructed distributedly using exhaustive search with respect to distances, i.e., Bellmann-Ford is run for sufficiently many iterations until all routes become stable. This approach has time complexity $\Omega(\text{SPD})$ and therefore cannot guarantee a running time of $\tilde{o}(n)$ on all graphs of diameter $\text{HD} \in \tilde{o}(n)$.

- sampling probabilities of $n^{-1/4}, n^{-3/8}, n^{-7/16}, \dots$, i.e., $p_i = n^{-(2^i-1)/2^{i+1}}$;
- expected set sizes of $\Theta(n^{3/4}), \Theta(n^{5/8}), \Theta(n^{9/16}), \dots$, i.e., $|S_i| \in \Theta(n^{1/2+1/2^i})$ w.h.p.;
- range parameters of $\Theta(n^{1/4} \log n), \Theta(n^{3/8} \log n), \Theta(n^{7/16} \log n)$, i.e., $h_i \in \Theta(n^{1/2-1/2^{i+1}} \log n)$;
- overlap parameters of $\Theta(n^{1/4} \log n), \Theta(n^{1/8} \log n), \Theta(n^{1/16} \log n)$, i.e., $\Delta_i \in \Theta(n^{1/2^{i+1}} \log n)$.

(Note that $L = \log \log n$ stages suffice to ensure that $S_L \in \Theta(\sqrt{n})$ w.h.p.) Running Algorithm BSP with parameters as above, we get that w.h.p., after $\tilde{O}(\sqrt{n})$ time, each node knows of the closest Δ_i nodes from S_{i-1} and how to route to them for all $1 \leq i \leq L$. But this is not sufficient: we also need to be able to route back from the nodes in S_i .

Given a node v , define $Y_i(v)$ to be the node closest to v in S_i (symmetry broken by identifiers), and let $C_v(i) \stackrel{\text{def}}{=} \{u \in V \mid Y_i(u) = v\}$, i.e., for each stage i , the sets $C_v(i)$ are a Voronoi decomposition of V with centers $Y_i(V)$. Note that routing from $Y_i(v)$ to $C_v(i)$ is not as simple as the other direction: While the depth of the tree rooted at $Y_i(v)$ is bounded by h_i , there is no non-trivial upper bound on the number of nodes in the tree. This can be solved by a number of standard techniques for tree routing (e.g., [27]). To minimize space consumption, we use the technique of [29], which constructs routing tables of size $\tilde{O}(1)$ and node labels of $\mathcal{O}(\log n)$ bits in $\mathcal{O}(h_i)$ time. In a nutshell, the idea is first to count the sizes of subtrees (which can be done in $\mathcal{O}(h_i)$ rounds) and then construct “mini routing tables” for the “heavy” part of the tree, where a node is considered heavy if its subtree contains at least $n/\lceil \sqrt{\log n} \rceil$ nodes. Then this process is applied recursively in the subtrees rooted at children of heavy nodes. From the description in [29], one can verify that each recursive step of the construction can be performed in time $\tilde{O}(h_i)$ in a tree of depth h_i in the **CONGEST**($\log n$) model. There are at most $\log_{\sqrt{\log n}} n$ recursive steps, summing up to a total of $\tilde{O}(h_i)$ rounds to construct labels and routing tables.

Formally, given natural numbers n and $L \leq \log \log n$, we define the following for $1 \leq i \leq L$.

- $p_0 \stackrel{\text{def}}{=} 1$, and $p_i \stackrel{\text{def}}{=} (\sqrt{n})^{-(2^L/(2^L-1))(2^i-1)/2^i}$.
- For each node v , $Y_v(i)$ is the node from S_i closest to v (ties broken by hop distance and ID).
- For each $u \in S_i$, define $C_u(i) \stackrel{\text{def}}{=} \{v \mid Y_v(i) = u\}$, and $C_u(0) \stackrel{\text{def}}{=} \{u\}$.
- For each node v , define $H_v(i) \stackrel{\text{def}}{=} \{u \in S_{i-1} \mid \text{wd}(v, u) \leq \text{wd}(v, Y_v(i))\}$.

Our construction maintains (w.h.p.) the following properties at stage $i \in \{1, \dots, L\}$.

- (1) S_i is a uniformly random subset of S_{i-1} , where $\Pr[v \in S_i] = p_i$ and $\Pr[v \in S_i \mid v \in S_{i-1}] = p_i/p_{i-1} = (\sqrt{n})^{-2^L/(2^i(2^L-1))}$.
- (2) For any node v , it is possible to route from v to $Y_v(i)$ on a least-weight path.
- (3) For any node v , it is possible to compute $Y_v(i)$ and $\text{wd}(v, Y_v(i))$ from the label of v .
- (4) For any node $u \in S_i$, it is possible to route from u to any node $w \in C_u(i)$ on a least-weight path.
- (5) For any node v , $H_v(i)$ is locally known at v , and it is possible to route from v to any node $u \in H_v(i)$ on a least-weight path (whose weight is known at v).

Suppose that we have such a hierarchy of L stages. Then, given the label of any node $w \in \bigcup_{1 \leq i \leq L} \bigcup_{u \in H_v(i)} C_u(i-1)$, node v can route a message to w as follows: First, find some $i \in \{1, \dots, L\}$ such that $w \in C_u(i-1)$ for some $u \in H_v(i)$ (cf. [Property \(3\)](#) and [Property \(5\)](#) of the construction).

The route from v to w is then defined by the concatenation of two shortest paths: the one from v to u , and the one from u to w (cf. [Property \(4\)](#) and [Property \(5\)](#)). Moreover, the long-range scheme will make sure that we can always route to any destination via the closest skeleton nodes in S_L , which is feasible due to [Property \(2\)](#) and [Property \(4\)](#). By always choosing from the available routes such that the weight of the computed route is minimal (which can be done by [Property \(3\)](#) and [Property \(5\)](#) for the short-range construction, and will also be possible for the long-range scheme), routing becomes stateless.

Stretch Analysis

We now bound the weight of the routes constructed by the stated scheme with respect to the weight of the shortest paths. We note that the argument for the general case is similar in spirit to the simple case of $i = 1$ illustrated in [Figure 2](#). We start with the following key lemma.

Lemma 4.6 *Suppose that for $v, w \in V$ and $1 \leq j \leq L$ we have that $w \notin \bigcup_{i=1}^j \bigcup_{u \in H_v(i)} C_u(i-1)$. Then (a) $\text{wd}(v, Y_v(j)) \leq (2j-1)\text{wd}(v, w)$, and (b) $\text{wd}(w, Y_w(j)) \leq 2j\text{wd}(v, w)$.*

Proof: We prove the lemma by induction on i , for a fixed j . More specifically, we show for each $1 \leq i \leq j$ that (a) $\text{wd}(v, v_i) \leq (2i-1)\text{wd}(v, w)$ and (b) $\text{wd}(w, w_i) \leq 2i\text{wd}(v, w)$. For the basis of the induction, consider $i = 0$ in Statement (b). In this case, since $S_0 = V$, we have that, $Y_w(0) = w$ and Statement (b) holds because $\text{wd}(v, w) \geq 0 = \text{wd}(w, w)$.

For the inductive step, assume that Statement (b) holds for $0 \leq i < j$ and consider $i+1$. Since trivially $w \in C_{Y_w(i)}(i-1)$, the premise of the lemma implies that $Y_w(i) \notin H_v(i+1)$. However, $Y_v(i+1) \in H_v(i+1)$, and hence we obtain

$$\begin{aligned} \text{wd}(v, Y_v(i+1)) &\leq \text{wd}(v, Y_w(i)) \\ &\leq \text{wd}(v, w) + \text{wd}(w, Y_w(i)) && \text{triangle inequality} \\ &\leq (2i+1)\text{wd}(v, w) && \text{by induction hypothesis} \end{aligned}$$

This proves part (a) of the claim. Using the above inequality we also obtain

$$\begin{aligned} \text{wd}(w, Y_w(i+1)) &\leq \text{wd}(w, Y_v(i+1)) && \text{wd}(w, Y_w(i+1)) \leq \text{wd}(w, u) \text{ for } u \in S_{i+1} \\ &\leq \text{wd}(w, v) + \text{wd}(v, Y_v(i+1)) && \text{triangle inequality} \\ &\leq (2i+2)\text{wd}(v, w) && \text{by the proof of part (a),} \end{aligned}$$

which proves part (b) of the claim, completing the inductive step. \blacksquare

[Lemma 4.6](#) allows us to prove the following positive result.

Corollary 4.7 *Let $v, w \in V$, and let $1 \leq i_0 \leq L$ be minimal such that $Y_w(i_0-1) \in H_v(i_0)$. Then $\text{wd}(v, Y_w(i_0-1)) + \text{wd}(Y_w(i_0-1), w) \leq (4i_0-3)\text{wd}(v, w) \in \mathcal{O}(L \cdot \text{wd}(v, w))$.*

Proof: Note that

$$\begin{aligned} \text{wd}(v, Y_w(i_0-1)) + \text{wd}(Y_w(i_0-1), w) &\leq \text{wd}(v, w) + 2\text{wd}(w, Y_w(i_0-1)) && \text{triangle inequality} \\ &\leq \text{wd}(v, w) + 4(i_0-1)\text{wd}(v, w) && \text{Lemma 4.6} \\ &= (4i_0-3)\text{wd}(v, w) \end{aligned}$$

and the corollary is proved. \blacksquare

On the other hand, if there is no i_0 as in the corollary, we can conclude from [Lemma 4.6](#) that routing via the skeleton nodes closest to source and destination, respectively, incurs bounded stretch.

Implementation and Time Complexity

We now explain how to construct the hierarchy efficiently in more detail, and analyze the time complexity of the construction. [Algorithm 2](#) gives the pseudocode of the above scheme. The algorithm is parametrized by the total number of nodes n and the number L of hierarchy stages. Appropriate constants c and c' are supposed to be predefined in accordance with the required lower bound on the probability of success.⁶

Algorithm 2: Distributed construction of data structure for close-distance routing at $v \in V$.	
input	$: n \in \mathbb{N}$ <i>// number of nodes</i>
	$L \in \{1, \dots, \log \log n\}$ <i>// number of stages in the hierarchy</i>
computes:	$l_v \in \{0, \dots, L\}$ <i>// level of v; $v \in S_i \Leftrightarrow l_v \geq i$</i>
	$\forall i \in \{1, \dots, L\} : Y_v(i) \in S_i$ <i>// closest node in S_i</i>
	$\forall i \in \{1, \dots, L\} : H_v(i) = \{w \in S_{i-1} \mid \text{wd}(v, w) \leq \text{wd}(v, Y_v(i))\}$
	$\forall i \in \{1, \dots, L\} \forall u \in H_v(i) : \text{next}_v(u), d_v(u)$ <i>// next routing hop (v if $v = u$) and distance to u</i>
1	for $i \in \{0, \dots, L\}$ do $p_i := (\sqrt{n})^{-(2^L/(2^L-1))(2^i-1)/2^i}$
2	$l_v := i$ with probability $p_i - \sum_{j=i+1}^L p_j$ for $i \in \{0, \dots, L\}$
3	for $i \in \{1, \dots, L\}$ do
4	$h_i := c \cdot \log n / p_i$ <i>// c and c' are predefined constants controlling the probability of failure</i>
5	$\Delta_i := c' \cdot h_i p_{i-1}$
6	if $l_v \geq i$ then $\text{source}(v) := (v, l_v)$ else $\text{source}(v) := \perp$
7	$(L_v(1), \dots, L_v(h_i)) := \text{BSP}(h_i, \Delta_i, \text{source})$ <i>// only $L_v(h_i)$ needed</i>
8	$H_v(i) := \emptyset$
9	repeat
10	let $(d, (u, l_u), w)$ be the next entry in $L_v(h_i)$ in ascending lexicographic order
11	$H_v(i) := H_v(i) \cup \{u\}$
12	$\text{next}_v(u) := w; d_v(u) := d$ <i>// exact shortest paths, no distinction of stages needed</i>
13	until $l_u \geq i;$
14	$Y_v(i) := u$ <i>// u is the node from S_{i+1} closest to v</i>
15	construct labels of stage i

Choosing the sets S_i is performed locally without communication. Each node v has level l_v chosen independently so that $\Pr[l_v \geq i] = p_i$ ([Line 2](#)). Setting $S_i \stackrel{\text{def}}{=} \{v \in V \mid l_v \geq i\}$ as indicated in the algorithm thus satisfies [Property \(1\)](#). In addition, the following properties are easily derived using the Chernoff bound, and we state them without proof.

Lemma 4.8 *For appropriate choices of the constants c, c' in [Algorithm 2](#), for all $1 \leq i \leq L$ it holds w.h.p. that:*

- $|S_i| \in \Theta(p_i n)$ ($|S_0| = n$).

⁶One can verify the properties of the construction and restart a failed iteration within $\mathcal{O}(\text{HD})$ time if desired, implying that the stretch guarantee becomes deterministic and the running time probabilistically bounded instead.

- For all $v \in V$, $|S_i \cap \text{ball}_v(h_i)| \in \Theta(\log n)$.
- For all $v \in V$, $H_v(i) \subset \text{ball}_v(h_i)$.
- For all $v \in V$, $|H_v(i)| \in \Theta(h_i p_{i-1}) = \Theta(p_{i-1} \log n / p_i)$.
- For all $v \in V$, $\Delta_i \geq |H_v(i)|$.

By these properties and [Theorem 4.4](#), [Property \(5\)](#) is satisfied w.h.p., because we invoke Algorithm BSP with sources S_i , depth parameter h_i , and overlap parameter Δ_i : after this invocation, each node v can identify the set $H_v(i)$ and route to any $u \in H_v(i)$ on a path of known weight; since $H_v(i) \subset \text{ball}_v(h_i)$ w.h.p., these routing paths are shortest paths. Moreover, the invocation of $\text{BSP}(h, \Delta, S_i)$ allows each node v also to learn what is $Y_v(i)$ and route to it on a shortest path of known weight, establishing [Property \(2\)](#). In order to satisfy [Property \(3\)](#), we simply add $Y_v(i)$ and $\text{wd}(v, Y_v(i)) = d_v(Y_v(i))$ to the label of v for all i . As discussed earlier, routing tables of size $\log^{\mathcal{O}(1)} n$ and labels of size $(1 + o(1)) \log n$ to route within $C_{Y_v(i)}$ can be constructed within $\tilde{\mathcal{O}}(h_i)$ rounds using the scheme from [\[29\]](#), and we add the respective tree label to v 's label to ensure [Property \(4\)](#).

We can therefore summarize the complexity of the construction as follows.

Lemma 4.9 *Given $1 \leq L \leq \log \log n$, constructing the L -stages short-range routing tables and labels can be done in $\mathcal{O}(L(\sqrt{n})^{2^L/(2^L-1)} \log^2 n) \subset \tilde{\mathcal{O}}((\sqrt{n})^{2^L/(2^L-1)})$ rounds, and the total label size of a node is $\mathcal{O}(L \log n)$.*

Proof: The implementation of stage $i \in \{1, \dots, L\}$ involves invoking BSP with parameters h_i and Δ_i , which, by [Theorem 4.4](#), takes

$$\mathcal{O}(h_i \Delta_i) = \mathcal{O}\left(\frac{p_{i-1} \log^2 n}{p_i^2}\right) = \mathcal{O}\left((\sqrt{n})^{2^L/(2^L-1)} \log^2 n\right)$$

rounds. In addition, we need to relabel the nodes, which, as explained above, can be done in time $\tilde{\mathcal{O}}(h_i) \subseteq \tilde{\mathcal{O}}(h_i \Delta_i)$, since the depth of the shortest paths tree is bounded by $h_i \leq h_i \Delta_i$. Since there are $L \leq \log \log n$ stages, the total number of rounds thus satisfies the stated bounds. With respect to the label size, note that each stage i adds to the label of node v the identifier of and distance to $Y_v(i)$ and a tree label of size $(1 + o(1)) \log n$, for a total of $\mathcal{O}(\log n)$ bits per stage. ■

4.3 Long-Distance Routing

We now explain how to route between the nodes in the top level of the hierarchy created by the short-range scheme. Our central concept is the *skeleton graph*, defined as follows.

Definition 4.10 (Skeleton Graph) *Let $G = (V, E, W)$ be a weighted graph. Given $S \subseteq V$ and $h \in \mathbb{N}$, the h -hop skeleton- S graph is the weighted graph $G_{S,h} = (S, E_{S,h}, W_{S,h})$ defined by*

- $E_{S,h} \stackrel{\text{def}}{=} \{\{v, w\} \mid v, w \in S, v \neq w, \text{ and } \text{hd}(v, w) \leq h\}$
- For $\{v, w\} \in E_{S,h}$, define $W_{S,h}(v, w)$ to be the h -weighted distance between v and w in G , i.e., $W_{S,h}(v, w) \stackrel{\text{def}}{=} \text{wd}_h(v, w)$.

The main idea in the long-distance scheme is to construct a skeleton graph with $S = S_L$ (the top level of the short-range hierarchy as constructed in [Section 4.2](#)). The choice of h needs to balance two goals: on the one hand, the skeleton graph needs to accurately reflect the distances of skeleton nodes

in G , and on the other hand, we must be able to quickly set up a tables that allow routing of small stretch between the skeleton nodes.

A simple but crucial observation on skeleton graphs is that if the skeleton S is a random set of nodes, and if $h \in \Omega(n \log n / |S|)$, then w.h.p., the distances in $G_{S,h}$ are equal to the corresponding distances in G . This means that it suffices to consider paths of $\mathcal{O}(n \log n / |S|)$ hops in G in order to find the exact distances in G . The following lemma formalizes this idea. (We state it for a skeleton *containing* a random subset; this generality will become useful in [Section 5.3](#).)

Lemma 4.11 *Let S_R be a set of random nodes defined by $\Pr[v \in S_R] = \pi$ independently for all nodes for some given π . Let $S \supseteq S_R$. If $\pi \geq c \log n / h$ for a sufficiently large constant $c > 0$, then w.h.p., $\text{wd}_{S,h}(v, w) = \text{wd}(v, w)$ for all $v, w \in S$.*

Proof: Fix $v, w \in S$. Clearly, $\text{wd}_{S,h}(v, w) \geq \text{wd}(v, w)$ because each path in $G_{S,h}$ corresponds to a path of the same weight in G . We need to show that $\text{wd}_{S,h}(v, w) \leq \text{wd}(v, w)$ as well. Let $p = \langle u_0 = v, u_1, \dots, u_{\ell(p)} = w \rangle$ be a shortest path connecting v and w in G , i.e., $W(p) = \text{wd}(v, w)$. We show, by induction on $\ell(p)$, that $\text{wd}(p) \leq \text{wd}_{S,h}(v, w)$ w.h.p.

For the basis of the induction note that if $\ell(p) \leq h$, then by definition $\text{wd}_{S,h}(v, w) \leq W(p) = \text{wd}(v, w)$ and we are done. For the inductive step, assume that the claim holds for all values of $\ell(p) \leq i$ for some $i \geq h$ and consider a path of length $\ell(p) = i + 1$. Now, $\mathbb{E}[|S \cap \{u_1, \dots, u_i\}|] \geq \mathbb{E}[|S_R \cap \{u_1, \dots, u_i\}|] = i\pi \geq h\pi \in \Omega(\log n)$, and hence, applying Chernoff's bound, we may conclude that w.h.p. the intersection is non-empty. Let $u \in \{u_1, \dots, u_i\} \cap S$. Since p is a shortest path in G , so are (v, \dots, u) and (u, \dots, w) . Both these paths are of length at most i , implying by the induction hypothesis that $\text{wd}_{S,h}(v, u) \leq \text{wd}(v, u)$ and $\text{wd}_{S,h}(u, w) \leq \text{wd}(u, w)$ w.h.p., respectively. Therefore $\text{wd}_{S,h}(v, w) \leq \text{wd}_{S,h}(v, u) + \text{wd}_{S,h}(u, w) \leq \text{wd}(v, u) + \text{wd}(u, w) = W(p) = \text{wd}(v, w)$, completing the induction. Note that the total number of events we consider throughout the induction is bounded by a polynomial in n , and since the probability of the bad events is polynomially small, the union bound allows us to deduce that the claim holds w.h.p. ■

Based on this observation, an obvious strategy to solve long-distance routing is to construct $G_{S,h}$ and compute its all-pairs shortest paths. But implementing this approach is not straightforward. First, the edges of the skeleton graph are virtual: each edge represents the shortest path of up to h hops in G ; and second, the number of skeleton graph edges may be as large as $\Omega(|S|^2)$. We solve both problems together: While computing the edges of the skeleton graph, we sparsify the graph, bringing the number of edges down to near-linear in the skeleton size. Once we are done, we can afford to let each skeleton node learn the full topology of the sparsified skeleton graph, from which approximate all-pairs routes and distances can be computed locally.

Technically, we use the classical concept of sparse spanners, defined as follows.

Definition 4.12 (Weighted k -Spanners) *Let $H = (V, E, W)$ be a weighted graph and let $k \geq 1$. A weighted k -spanner of H is a weighted graph $H' = (V, E', W')$ where $E' \subseteq E$, $W'(e) = W(e)$ for all $e \in E'$, and $\text{wd}_{H'}(u, v) \leq k \cdot \text{wd}_H(u, v)$ for all $u, v \in V$ (where wd_H and $\text{wd}_{H'}$ denote weighted distances in H and H' , respectively).*

We shall compute a spanner of the skeleton graph, while running on the underlying physical graph, without ever constructing the skeleton graph explicitly. We do this by simulating the spanner construction algorithm of Baswana and Sen [3] on the implicit skeleton graph. Let us recall the algorithm of [3]; we use a slightly simpler variant that may select some additional edges, albeit without affecting the probabilistic upper bound on the number of spanner edges (cf. Lemma 4.14). The input is a graph $H = (V_H, E_H, W_H)$ and a parameter $k \in \mathbb{N}$.

-
1. Initially, each node is a singleton *cluster*: $R_1 := \{\{v\} \mid v \in V_H\}$.
 2. Repeat $k - 1$ times (the i^{th} iteration is called “phase i ”):
 - (a) Each cluster from R_i is *marked* independently with probability $|V_H|^{-1/k}$. R_{i+1} is defined to be the set of clusters marked in phase i .
 - (b) If v is a node in an unmarked cluster:
 - i. Define Q_v to be the set of edges that consists of the lightest edge from v to each of the clusters $v \in R_i$ is adjacent to.
 - ii. If v has no adjacent marked cluster, then v adds to the spanner all edges in Q_v .
 - iii. Otherwise, let u be the closest neighbor of v in a marked cluster. In this case v adds to the spanner the edge $\{v, u\}$, and also all edges $\{v, w\} \in Q_v$ with $(W_H(v, w), w) < (W_H(v, u), u)$ (i.e., the identifiers w, u break symmetry in case $W_H(v, w) = W_H(v, u)$). Furthermore v *joins* the cluster of u (i.e., if u is in cluster X , then $X := X \cup \{v\}$).
 3. Each node v adds, for each cluster $X \in R_k$ it is adjacent to, the lightest edge connecting it to X .
-

For this algorithm, Baswana and Sen prove the following result.

Theorem 4.13 ([3]) *Given a weighted graph $H = (V_H, E_H, W_H)$ and an integer $k \geq 1$, the algorithm above computes a $(2k - 1)$ -spanner of the graph. It has $\mathcal{O}(k|V_H|^{1+1/k} \log n)$ edges w.h.p.⁷*

Constructing the Skeleton Graph

In our case, each edge considered in Steps (2b) and (3) of the spanner algorithm corresponds to a shortest path. Essentially, we implement these steps in our setting by letting each skeleton node find its closest $\mathcal{O}(|S|^{1/k} \log n)$ clusters (w.h.p.) by running Algorithm BSP. We now explain how. First, all nodes v in a cluster X use the same source identifier $\text{source}(v) = X$ (as if they were connected by a 0-weight edge to a virtual node X). This ensures that the overlap parameter needs to account for the number of detected *clusters* only, i.e., the number of nodes per cluster is immaterial. Note that this implies that the plain version of Algorithm BSP thus will not permit to determine to which node a skeleton edge connects; hence we append to each communicated triple (d, s, next) the identifier of the actual endpoint $u \in SN(s)$ of the respective path and store it when adding a corresponding triple to L_v (without otherwise affecting the algorithm). We refer to the modified algorithm as BSP' . Second, regarding the range parameter, Lemma 4.11 shows that it is sufficient to consider paths of $\mathcal{O}(n \log n / |S|)$ hops only. Finally, the following lemma implies that we may modify the spanner construction algorithm in a way that allows us to use a small overlap parameter.

⁷In [3], it is proved that the expected number of edges is $\mathcal{O}(k|V_H|^{1+1/k})$. The modified bound directly follows from Lemma 4.14.

Lemma 4.14 *W.h.p., the execution of the centralized spanner construction algorithm yields identical results if in Steps (2b) and (3), each node considers the lightest edges to the $c|V_H|^{1/k} \log n$ closest clusters only (for a sufficiently large constant $c > 0$).*

Proof: Fix a node v and a phase $1 \leq i < k$. If v has at most $c|V_H|^{1/k} \log n$ adjacent clusters, the lemma is trivially true. So suppose that v has more than $c|V_H|^{1/k} \log n$ adjacent clusters. By the specification of Step (2b), we are interested only in the clusters closer than the closest marked cluster. Now, the probability that none of the closest $c|V_H|^{1/k} \log n$ clusters is marked is $(1 - |V_H|^{-1/k})^{c|V_H|^{1/k} \log n} \in n^{-\Omega(c)}$. In other words, choosing a sufficiently large constant c , we are guaranteed that w.h.p., at least one of the closest $c|V_H|^{1/k} \log n$ clusters is marked. Regarding Step (3), observe that a cluster gets marked in all of the first $k - 1$ iterations with independent probability $|V_H|^{-(k-1)/k}$. By Chernoff's bound, the probability that more than $c|V_H|^{1/k} \log n$ clusters remain in the last iteration is thus bounded by $2^{-\Omega(c \log n)} = n^{-\Omega(c)}$. Therefore, w.h.p. no node is adjacent to more than $c|V_H|^{1/k} \log n$ clusters in Step (3), and we are done. ■

As a consequence of Lemma 4.14, we may invoke Algorithm BSP with $\Delta \in \Theta(|S|^{1/k} \log n)$, and the time complexity of the invocation is $\mathcal{O}(|V| \cdot |S|^{-1+1/k} \log^2 n)$. Detailed pseudo-code of our implementation is given in Algorithm 3. Each skeleton node $v \in S$ records the ID of its cluster in phase i as $F_i(v)$; nodes in $V \setminus S$ or those who do not join a cluster in some round i have $F_i(v) = \perp$. Algorithm 4 is used as subroutine to implement Steps (2b) or (3) (Lines 9 or 20 of Algorithm 3, respectively).

To prove the algorithm correct, we show that its executions can be mapped to executions of the centralized algorithm, and then apply Theorem 4.13. Below, we sketch the main points of such a mapping. The implementation of Algorithm 3 is quite straightforward. Note that the broadcast steps in Line 1, 8, and 19 ensure that all nodes know the clusters and which are the active clusters in each phase. The random choices (Line 7) are made by cluster leaders, namely the nodes v for which $F_i(v) = v$. Lines 10–18 are local computations each node does to get a global picture of the clusters for the next phase. The correctness of the implementation of the edge selection of Steps 2b and 3 of the centralized algorithm by Algorithm 4 was discussed above. We summarize with the following lemma.

Lemma 4.15 *Suppose the set S input to Algorithm 3 contains a uniformly random subset S_R of V and set $h(S_R) \stackrel{\text{def}}{=} c \cdot n \log n / |S_R|$ for a sufficiently large constant c . Then w.h.p. the following holds.*

- (i) *Algorithm 3 computes a weighted $(2k - 1)$ -spanner of the skeleton graph $G_{S, h(S_R)}$ that is known at all nodes and has $\mathcal{O}(|S|^{1+1/k} \log n)$ edges.*
- (ii) *The weighted distances between nodes in S are identical in $G_{S, h(S_R)}$ and G .*
- (iii) *The algorithm terminates in $\tilde{\mathcal{O}}(n/|S_R|^{1-1/k} + |S|^{1+1/k} + \text{HD})$ rounds.*

Proof: To prove Statement (i), we note that Algorithm 3 simulates the centralized algorithm, except for considering only the closest $\mathcal{O}(|S|^{1/k} \log n)$ clusters in Lines 9 and 20. By Lemma 4.14 and by Theorem 4.4, this results in a (simulated) correct execution of the centralized algorithm w.h.p. Hence Statement (i) follows from Theorem 4.13.

Regarding Statement (ii), observe that if $h(S_R) \geq n - 1$, the statement holds by definition since shortest paths cannot contain cycles and thus $G_{S, h(S_R)} = G_{S, n-1}$. Otherwise, we have that $|S_R| \geq c \cdot \log n$, implying by Chernoff's bound that w.h.p., the probability to select a node into S is $\pi \in$

Algorithm 3: Construction of long range routing skeleton at $v \in V$.

```

input  :  $S$ : set of skeleton nodes
           $k$ : integer in  $[1, \log n]$            //determines approximation ratio and number of spanner edges
output:  $E_{h,k}$ : spanner edges of skeleton graph           //h is defined in Line 4
           $W_{h,k} : E_{h,k} \rightarrow \mathbb{R}^+$            //weights of spanner edges
1  $R_1 := \{\{w\} \mid w \in S\}$            //initial clusters are singletons of S
2 Broadcast  $R_1$  to all nodes
3 foreach  $w \in V$  do if  $w \in R_1$  then  $F_1(w) := w$  else  $F_1(w) := \perp$            //initializing leaders
4  $h := c \cdot n \log n / |S|$            //the constant c controls the probability of failure
5  $\Delta := c \cdot |S|^{1/k} \log n$ 
6 for  $i := 1$  to  $k - 1$  do
7    $R_{i+1} :=$  uniformly random subset of  $R_i$  of size  $|S|^{1-i/k} = |R_i|/|S|^{1/k}$            //select marked clusters
8   Broadcast  $R_{i+1}$  to all nodes
9    $(E(i), W(i)) := \text{edges}(F_i, R_{i+1}, h, \Delta)$            //select spanner edges, phase i
10  foreach  $w \in V$  do
11    if  $F_i(w) \in R_{i+1}$  then
12       $F_{i+1}(w) := F_i(w)$ 
13    else
14      Let  $E_w$  be the edges incident to  $w$  in  $E(i)$ 
15      if  $E_w \neq \emptyset$  then
16        Let  $\{w, u\}$  be the heaviest edge in  $E_w$ 
17        if  $\text{marked}(u)$  then  $F_{i+1}(w) := F_i(u)$ 
18      else  $F_{i+1}(w) := \perp$ 
19    Broadcast  $F_{i+1}$  to all nodes
20  $(E(k), W(k)) := \text{edges}(F_k, \emptyset, h, \Delta)$            //final phase
21 foreach  $e \in \bigcup_{i=1}^k E(i)$  do  $W_{h,k}(e) := W(k)(e)$ 
22 Broadcast  $E_{h,k} := \bigcup_{i=1}^k E(i)$  and  $W_{h,k}$  to all nodes

```

$\Theta(|S_R|/n) = \Theta((c \log n)/h(S_R))$. As by assumption c is sufficiently large, Statement (ii) now follows from Lemma 4.11.

For Statement (iii), consider first an invocation of Algorithm 4. By Theorem 4.4, the invocation of Algorithm 1 in Line 3 takes $\mathcal{O}(\Delta h) \subset \tilde{\mathcal{O}}(|S|^{1/k} h(S_R)) = \tilde{\mathcal{O}}(n|S|^{1/k}/|S_R|)$ rounds. The broadcast of Line 10 is done globally. Each skeleton node may communicate up to $\mathcal{O}(|S|^{1/k} \log n)$ pieces of information for a total of $\tilde{\mathcal{O}}(|S|^{1+1/k})$ items. Doing this over a global BFS tree takes $\tilde{\mathcal{O}}(\text{HD} + |S|^{1+1/k})$ rounds. As $k \leq \log n$, the total cost of all invocations of Algorithm 4 is thus bounded by $\tilde{\mathcal{O}}(n|S|^{1/k}/|S_R| + |S|^{1+1/k} + \text{HD})$ rounds. Consider now Algorithm 3. The only non-local steps other than the invocations of Algorithm 4 are the broadcasts, of which the most time consuming is the one in Line 22, which takes $\mathcal{O}(k|S|^{1+1/k} \log n + \text{HD}) \subset \tilde{\mathcal{O}}(|S|^{1+1/k} + \text{HD})$ rounds. ■

Routing on the Skeleton Graph

Algorithm 3 constructs a $(2k - 1)$ -spanner of the skeleton graph and made it known to all nodes. This enables each node to determine low-stretch routing paths between any two skeleton nodes in

Algorithm 4: edges: Edge detection and announcement for long range routing skeleton at $v \in V$.

input : $F : V \rightarrow V \cup \{\perp\}$ //locally known, v 's leader if v is in a cluster, otherwise \perp
 $R \subseteq V$ //globally known, indicates (identifiers of leaders of) marked clusters
 h //globally known, depth parameter of the search
 Δ //globally known, number of closest source clusters to detect

output: E_+ : edges added to the spanner
 $W_+ : E_+ \rightarrow \mathbb{R}^+$ edge weights

```

1 foreach  $w \in V$  do
2    $\text{source}(w) := \begin{cases} (F(v), 0) & \text{if } F(v) \notin R \cup \{\perp\} \\ (F(v), 1) & \text{if } F(v) \in R \\ \perp & \text{else} \end{cases}$  //distinguish marked from unmarked clusters
3  $L_v := \text{BSP}'(h, \Delta, \text{source})$  //variant of Algorithm BSP that keeps track of path endpoints
4  $E_+ := \emptyset$ 
5 if  $F_v \notin R \cup \{\perp\}$  then
6   //for each entry  $(\text{wd}, (f, b), u, w) \in L_v$ ,  $u$  is the next hop on a path of weight  $\text{wd}$  to  $w$  in cluster  $f$ 
7    $L_v := L_v \setminus \{(0, (F(v), 0), v, v)\}$  //remove loops (clusters are in distance 0 of themselves)
8   //recall that  $L_v$  is ordered; first entry with  $b = 1$  corresponds to closest marked cluster
9   foreach  $(\text{wd}, (f, b), u, w) \in L_v$  do
10    broadcast  $(\text{wd}, \{v, w\})$  //all nodes perform operation!
11     $E_+ := E_+ \cup \{v, w\}$ 
12     $W_+(\{v, w\}) := \text{wd}$ 
13    if  $f \in R$  then
14      break //f is closest marked cluster
15 return  $(E_+, W_+)$ 

```

$G_{S, h(S_R)}$ by local computation. To use this information, we must be able, for each spanner edge $\{s, t\} \in E_{S, h(S_R)}$, to route on a corresponding path in G , i.e., a path of weight $W_{S, h(S_R)}(s, t)$. Since we rely on Algorithm BSP during the construction of the spanner, Theorem 4.4 shows that we can use the computation to enable for each such edge to route from s to t or from t to s : if, say, s added the edge to the spanner, then following the pointers computed during the execution of Algorithm BSP yields a path of weight $W_{S, h(S_R)}(s, t)$ from s to t . However, in this case t might not add $\{s, t\}$ to the spanner as well, and hence there is no guarantee that we have sufficient information to route in *both* directions.⁸ To resolve this issue, we add a post-processing step where we “reverse” the unidirectional routing paths, i.e., inform the nodes on the paths about their predecessors. Note that this cannot be done in a purely local manner, as exchanging the Bellmann-Ford routing pointers between neighbors will not tell a node $s \in S$ which pointer to follow to reach a specific node $t \in S$ for which $\{s, t\}$ is part of the spanner. However, Corollary 4.4 states that the (unidirectional) routing paths at our disposal have at most $h(S_R)$ hops. Taking into account that the spanner has few edges, it follows that establishing bidirectional routing pointers can be performed sufficiently fast.

Lemma 4.16 *Let $\{s, t\}$ be an edge of the spanner $G_{S, h(S_R)}$ that is selected by Algorithm 3. W.h.p.,*

⁸Note that unidirectionality is not an artifact of the specific implementation we picked. E.g., in a star graph, the center has degree $n - 1$, as it does in the spanner. Hence we cannot expect the Bellmann-Ford pointers to give sufficient information for bidirectional routing without further processing.

after completing the algorithm, each node on the least-weight s - t path of at most $h(S_R)$ hops in G determine the next hop on this path and the weight of the remaining subpath when routing from s to t within $\tilde{O}(n/|S_R| + |S|^{1+1/k})$ rounds

Proof: For each edge $\{s, t\}$ added to the spanner by a node s , we route a message on the shortest path of at most $h(S_R)$ hops from s to t in G . This message initially contains the weight of the path, and each node on the path subtracts the weight of the incoming path from this value. By [Theorem 4.4](#) this is feasible. When a node receives the message, it records the immediate sender as the next hop on the path to s and the weight for future reference. By [Lemma 4.15](#), there are at most $\tilde{O}(|S|^{1+1/k})$ edges in the constructed spanner of $G_{S, h(S_R)}$ w.h.p., implying that the maximal number of messages routed over each edge of G is bounded by $\tilde{O}(|S|^{1+1/k})$ w.h.p. as well. Moreover, no routing path has more than $h(S_R) = c \cdot n \log n / |S_R|$ hops. Since the messages traverse shortest h -hop paths, all of them reach their destinations within the stated number of rounds [\[18\]](#). ■

We now summarize the properties of the long-distance scheme.

Theorem 4.17 *Suppose the set S input to [Algorithm 3](#) is a superset of a uniformly random subset $S_R \subseteq V$ and $k \in \{1, \dots, \log n\}$. Then, w.h.p., within $\tilde{O}(n|S|^{1/k}/|S_R| + |S|^{1+1/k} + \text{HD})$ rounds, there are routing tables for routing between nodes in S with stretch $(2k - 1)$.*

Proof: Directly follows from [Lemmas 4.15](#) and [4.16](#). ■

4.4 Putting the Pieces Together

Equipped with the results for the short-range and for the long-distance routing, we can state the overall algorithm as a simple composition of the two, linked by identifying the skeleton set from the long-distance algorithm with the top level of the hierarchy S_L of the short-range algorithm. We run the long-range algorithm with parameter k to construct and make globally known the routing skeleton and apply the short-range routing scheme with parameter L to deal with nearby nodes.

Recall that the label of a node w is $\lambda(w) = \langle Y_w(i), \text{wd}(w, Y_w(i)), \text{tree}_w(i) \rangle_{i=0}^L$, where $\text{tree}_w(i)$ denotes the label of v in the tree on $C_{Y_w(i)}$, and $Y_w(0)$ is simply w . Given the label $\lambda(w)$ to a node v , v decides on the next routing hop as follows.

- If $Y_v(i) = Y_w(i)$ for some i , choose the next routing hop within $C_{Y_v(i)}$ to w according to the respective tree label. In this case, d is the distance from v to w in the tree (which can be computed from the distances of v and w to the root $Y_v(i)$ and whether the next routing hop is the parent of v or a child).
- Otherwise, node v determines for each $i \in \{1, \dots, L\}$ whether $Y_w(i-1) \in H_v(i)$. If so, it computes $d_i \stackrel{\text{def}}{=} \text{wd}(v, Y_w(i-1)) + \text{wd}(Y_w(i-1), w)$. Otherwise set $d_i \stackrel{\text{def}}{=} \infty$.
- Next, denote by $S_v \subseteq S_L$ the set of skeleton nodes v for which it stores a routing pointer and the corresponding path weight, and let for $s \in S_v$ d_s be this weight. We define wd^k to be the distance function on the spanner of the skeleton graph. Node v computes $d_{L+1} \stackrel{\text{def}}{=} \min_{s \in S_v} \{d_s + \text{wd}^k(s, Y_w(L)) + \text{wd}(Y_w(L), w)\}$.
- Finally, v computes $d \stackrel{\text{def}}{=} \min_{i \in \{1, \dots, L+1\}} \{d_i\}$, and determines the next routing hop in accordance with the corresponding path (ties broken by preferring smaller i), where we use the routing

mechanism from [Corollary 4.5](#).

Since v stores the tree routing tables for all trees on $C_{Y_v(i)}$, the sets H_i and the distances to the nodes in H_i , and the complete spanner of the skeleton graph, together with the label $\lambda(w)$ it has the necessary information to perform all the above computations. Moreover, a next routing hop is always determined, since $Y_v(L) \in S_v$ (by [Property \(2\)](#) of the short-range scheme) and therefore the set of considered paths in the second step is non-empty. Finally, the routing decision is stateless, since it depends on the local routing tables of v and $\lambda(w)$ only.

In order to show that indeed a route to w of bounded stretch is determined by the above routing decisions, we will show two properties. First, the value d computed is the weight of a path of bounded stretch whose next routing hop u is exactly the one computed by v , and second, the next node u on the path will compute a distance of at most $d - \text{wd}(v, u)$ to w . Since edge weights are strictly positive, the latter immediately implies that the routes are acyclic and will eventually reach their destination.

Lemma 4.18 *Fix any choice of the parameters L and k of the short range and long distance schemes, respectively. For any node v and label $\lambda(w)$, consider the distance value d and next routing hop next computed by v according to the above scheme. Then w.h.p., $d \leq (8kL - 1)\text{wd}(v, w)$ and next will compute a value $d' \leq d - W(v, \text{next})$.*

Proof: We show that $d \leq (8kL - 1)\text{wd}(v, w)$ first. If $w \in \bigcup_{u \in H_v(i)} C_u(i - 1)$ for some $i \in \{1, \dots, L\}$, observe that $d \leq \text{wd}(v, Y_w(i)) + \text{wd}(Y_w(i), w)$, and thus by [Corollary 4.7](#) $d \leq (4L - 3)\text{wd}(v, w)$. Otherwise, we have that $d = d_{L+1}$ since no other routes are known to v . By definition and [Property \(2\)](#) of the short-range scheme, $d_{L+1} \leq \text{wd}(v, Y_v(L)) + \text{wd}^k(Y_v(L), Y_w(L)) + \text{wd}(Y_w(L), w)$. We bound

$$\begin{aligned}
& \text{wd}(v, Y_L(v)) + \text{wd}^k(Y_v(L), Y_w(L)) + \text{wd}(Y_L(w), w) \\
& \leq \text{wd}(v, Y_L(v)) + (2k - 1)\text{wd}(Y_L(v), Y_L(w)) + \text{wd}(Y_L(w), w) && \text{by Theorem 4.17} \\
& \leq 2k\text{wd}(Y_L(v), v) + (2k - 1)\text{wd}(v, w) + 2k\text{wd}(w, Y_L(w)) && \text{triangle inequality} \\
& \leq (2k(4L - 1) + (2k - 1))\text{wd}(v, w) && \text{Lemma 4.6} \\
& = (8kL - 1)\text{wd}(v, w),
\end{aligned}$$

proving that indeed $d \leq (8kL - 1)\text{wd}(v, w)$.

Now let next be the routing hop corresponding to d computed by v . Due to [Property \(4\)](#) and [Property \(5\)](#) of the short-range scheme, there are the following three cases:

- next is on the shortest path from v to $Y_w(i - 1) \in H_v(i)$ for some $i \in \{1, \dots, L\}$ (this covers also the case that $Y_v(i - 1) = Y_w(i - 1)$, and in the tree on $C_{Y_w(i-1)}$ the connecting path traverses the root $Y_w(i - 1)$);
- next is on a path of weight d_s to the node $s \in S_v$ minimizing the expression $d_s + \text{wd}^k(s, Y_w(L)) + \text{wd}(Y_w(L), w)$;
- next is on the shortest path from $Y_w(i)$ to w for some $i \in \{1, \dots, L\}$ (i.e., $Y_v(i) = Y_w(i)$, and in the tree on $C_{Y_w(i)}$ the connecting path does not traverse the root $Y_w(i)$).

Regarding the first case, observe that since we are talking about shortest paths in G (not shortest h -hop paths), any source closer to next than $Y_w(i - 1)$ will also be closer to v than $Y_w(i - 1)$. Hence $Y_w(i - 1) \in H_{\text{next}}(i)$. Since $\text{wd}(\text{next}, Y_w(i - 1)) = \text{wd}(v, Y_w(i - 1)) - W(v, \text{next})$, consequently next will compute a distance of at most $d - W(v, \text{next})$ to w .

In the second case, next is either the next hop on a routing path as constructed in [Corollary 4.5](#) or

as constructed by the “path reversal” from [Lemma 4.16](#). Either way, the statements show that next will know the next routing hop to s as well as the weight of the path. Since it knows the entire skeleton graph, it will thus compute a distance of at most $d_s - W(v, \text{next}) + \text{wd}^k(s, Y_w(L)) + \text{wd}(Y_w(L), w)$ to w as claimed.

For the third and final case, the statement trivially holds, since routing in $C_{Y_w(i)}$ according to the tree routing table is on shortest paths and will clearly lead to another node in $C_{Y_w(i)}$. ■

It is fairly straightforward to set k and L to obtain a trade-off between the stretch of the routing scheme and the construction time. Specifically, we can now state our main result as follows.

Theorem 4.19 *Let $1/2 \leq \alpha \leq 1$ be given. Define $k \stackrel{\text{def}}{=} \lceil 1/(2\alpha - 1) \rceil$ if $\alpha \geq 1/2 + 1/\log n$, and $k \stackrel{\text{def}}{=} \log n$ otherwise. Tables for stateless routing and distance approximation with stretch $\rho(\alpha) = 8k \lceil \log(k+1) \rceil - 1$ and label size $\mathcal{O}(\log(k+1) \log n)$ can be constructed in the **CONGEST** model in $\tilde{\mathcal{O}}(n^\alpha + \text{HD})$ rounds. In particular, $\rho(1/2) \in \mathcal{O}(\log n \log \log n)$ and $\rho(\alpha) \in \mathcal{O}(1)$ for any constant choice of $\alpha > 1/2$.*

Proof: A stretch bound of $8kL - 1$ and the fact that the destination will indeed be reached when following the computed pointers follows from [Lemma 4.18](#). By [Lemma 4.9](#), the running time of the short-range construction is bounded by $\tilde{\mathcal{O}}((\sqrt{n})^{2^L/(2^L-1)})$ rounds w.h.p. The time required for the skeleton construction is, by [Theorem 4.17](#), $\tilde{\mathcal{O}}(n/|S_L|^{1-1/k} + |S_L|^{1+1/k} + \text{HD}) = \tilde{\mathcal{O}}((\sqrt{n})^{1+1/k} + \text{HD})$ w.h.p. To match the desired running time bound of $\tilde{\mathcal{O}}(n^\alpha + \text{HD})$ rounds, it thus suffices that $\max\{1/k, 1/(2^L - 1)\} \geq 2\alpha - 1 - 1/\log n$ (an additive $1/\log n$ in the exponent maps to a constant factor). By choice of k , this inequality holds for $L \stackrel{\text{def}}{=} \lceil \log(k+1) \rceil$. The stretch is thus bounded by

$$\rho(\alpha) + 1 \leq 8kL = \begin{cases} 8\lceil 1/(2\alpha - 1) \rceil \lceil \log \lceil 1/(2\alpha - 1) \rceil \rceil + 1 & \text{for } \alpha > 1/2 \\ 8 \log n \lceil \log \log n + 1 \rceil & \text{for } \alpha = 1/2. \end{cases}$$

The bound on the label size follows from [Lemma 4.9](#), our choice of L , and the fact that the long-distance scheme adds only $\mathcal{O}(\log n)$ bits to the label. ■

The space complexity of our scheme, i.e., the number of bits of the computed routing tables, is also straightforward to bound.

Corollary 4.20 *The size of the routing table at node v computed by the algorithm referenced in [Theorem 4.19](#) is $\tilde{\mathcal{O}}(n^\alpha)$.*

Proof: Observe that the dominant terms in memory consumption are (i) storing the sets $H_v(i)$ and the next pointers to them for the short-range routing scheme, (ii) storing the routing information for the paths from the roots $u \in S_i$ of the trees induced by the sets $C_u(i)$, and (iii) storing $G_{S_L, h(S_L)}$ and the next pointers for the long-range scheme. Trivially, the encoding of $G_{S_L, h(S_L)}$ cannot require more than $\tilde{\mathcal{O}}(n^\alpha)$ memory, as it is broadcasted globally over the BFS tree. The routing information from $C_u(i)$ to the nodes in its tree is $\log^{\mathcal{O}(1)} n$ bits [29]. The term from (i) originates from calls to Algorithm BSP. The routing information that needs to be stored consists of the history of the list maintained by Algorithm BSP. Hence, if such a call has depth and overlap parameters h and Δ , the memory required is $\mathcal{O}(h\Delta \log n)$. Hence the memory bound for (i) directly follows from the running time bound from [Lemma 4.9](#). ■

5 Extensions and Applications

5.1 Distance Sketches

The problem of distributed distance sketches requires each node to have a label and store a small amount of information (called the *sketch*), so that each node v can estimate the distance to each other node u when given the label of u .⁹ Technically speaking, we already solved this problem, since our machinery enables to estimate distances with small stretch. However, since $\alpha \geq 1/2$, the basic construction will always consume $\tilde{\Omega}(\sqrt{n})$ memory.

If we discard the routing information, we can reduce the space requirements of the sketches at the expense of also increasing the stretch. To this end, we need to reduce the maximal size of the sets $H_v(i)$ as well as the space consumed for storing information on the skeleton graph. Our idea is as follows. First, we change the sampling probabilities of the sets S_i so that $|S_i| \in \Theta(n^{1-i/L})$, where i ranges from 0 to L . With this choice, the expected number of nodes from S_i that are closer to a given node than the closest node from S_{i+1} is $\Theta(n^{1/L})$, implying that $\mathbb{E}[|H_v(i)|] \in \Theta(n^{1/L})$ for all i . Second, we do not choose S_L as skeleton set, but rather S_{i_0} for $i_0 \approx L/2$, so that the skeleton can be constructed quickly. To continue applying the short-range scheme beyond stage i_0 without increasing the asymptotic time complexity of the construction, we construct temporary distance sketches and labels for the skeleton (using the long-range scheme with skeleton set S_{i_0}), which allows nodes to estimate their distance to skeleton nodes locally. Ensuring that the S_i are subsets of the skeleton for $i \geq i_0$, each node thus can simulate the short-range algorithm's detection of the sets H_i and the respective distance computation locally, based on its estimated distance to skeleton nodes; the price we pay is increasing the stretch by factor $\mathcal{O}(L)$ due to imprecise distances.

Theorem 5.1 *Given any integer $k \in [1, \dots, \log n]$, distance sketches with stretch $\rho(k) = 2k(8k-3) \in \mathcal{O}(k^2)$, label size $\mathcal{O}(k \log n)$, and sketch size $\tilde{\mathcal{O}}(n^{1/(2k)})$ can be constructed w.h.p. in the **CONGEST** model in $\tilde{\mathcal{O}}(n^{1/2+1/(2k)} + \text{HD})$ rounds.*

Proof: We use the following algorithm, parametrized by k .

1. Run the short-range with k stages and expected set sizes $\mathbb{E}[|S_i|] = n^{i/(2k)}$ for $i \in \{0, \dots, k\}$.
2. Run the long-range scheme on the skeleton S_k .
3. For $k+1 \leq i \leq 2k-1$, sample set $S_i \subseteq S_{i-1}$, where each node is picked uniformly with probability $n^{-1/(2k)}$ in each step. Each node in S_k broadcasts its membership information.
4. For each pair $v \in V$ and $s \in S_k$, set $\text{wd}'(v, s) \stackrel{\text{def}}{=} \text{wd}(v, Y_v(L)) + \text{wd}(Y_v(L), s)$. For $k+1 \leq i \leq 2k-1$, compute at each node v the closest node $Y_v(i) \in S_i$ w.r.t. wd' , and the set $H_v(i) := \{s \in S_{i-1} \mid \text{wd}'(v, s) \leq \text{wd}'(v, Y_v(i))\}$. Set $H_v(2k) \stackrel{\text{def}}{=} S_{2k-1}$.
5. Store at each node v , for each $i \in \{1, \dots, 2k\}$: (i) the set $H_v(i)$; (ii) for each $u \in H_v(i)$, the value $\text{wd}(v, u)$ if $i \leq k$, or $\text{wd}'(v, u)$ if $i > k$. Label node v by $\lambda(v) = (Y_v(i), d(i))_{i \in \{0, \dots, 2k-1\}}$, where $d(i) \stackrel{\text{def}}{=} \text{wd}(v, Y_v(i))$ if $i < k$, and $d(i) \stackrel{\text{def}}{=} \text{wd}'(v, Y_v(i))$ otherwise.

Given label $\lambda(w)$ (which is clearly of size $\mathcal{O}(k \log n)$), node v estimates the distance to w by finding

⁹The formulation in [6] permits to use *both* sketches to approximate the distance. However, from the distributed point of view it is more appropriate to assume that only a minimal amount of information is exchanged.

the smallest $i \in \{1, \dots, 2k\}$ so that $Y_w(i-1) \in H_v(i)$ and adding the respective distance estimates from v to $Y_w(i-1)$ (locally known) and from $Y_w(i-1)$ to w (from the label). Such an i always exists because $H_v(2k) = S_{2k-1}$. This completes the description of the algorithm.

By [Corollary 4.7](#), the stretch of the routes represented by the labels and sketches would be $8k-3 \in \mathcal{O}(k)$ w.h.p. if all distances were exact. The approximation ratio is obtained by multiplying this value by the maximal stretch of any distance estimates employed in the construction. Up to stage k , all values are exact w.h.p. Thereafter, we use estimates of distances between skeleton nodes and all other nodes. By the triangle inequality, we have for all $v \in V$ and $s \in S$ that

$$\text{wd}(v, s) \leq \text{wd}(v, Y_v(i_0)) + \text{wd}(Y_v(i_0), s) \leq \text{wd}(v, Y_v(i_0)) + \text{wd}^k(Y_v(i_0), s) = \tilde{\text{wd}}(v, s).$$

On the other hand,

$$\begin{aligned} \text{wd}(v, Y_v(i_0)) + \text{wd}^k(Y_v(i_0), s) &\leq \text{wd}(v, s) + \text{wd}^k(Y_v(i_0), s) && \text{by definition of } Y_v(i_0) \\ &\leq 2k \cdot \text{wd}(Y_v(i_0), s) && \text{by Theorem 4.17.} \end{aligned}$$

Hence the stretch of the distance estimates is bounded by $\rho(k) = 2k(8k-3) \in \mathcal{O}(k^2)$.

By Chernoff's bound, for all i we have that $|S_i| \in \Theta(n^{(2k-i)/(2k)})$ w.h.p. Hence, the non-local part of the construction can be performed with overlap parameter $\Delta_i \in \Theta(n^{1/(2k)} \log n)$ and distance parameter $h_i \in \tilde{\Theta}(n^{i/(2k)})$ for all $i \in \{1, \dots, k\}$. We conclude the claimed running time and memory bounds of $\tilde{\mathcal{O}}(n^{1/2+1/(2k)})$ (time $\tilde{\mathcal{O}}(h_i \Delta_i)$ for each step of the short-range scheme, time $\tilde{\mathcal{O}}(n/|S_k|^{1-1/k} + |S_k|^{1+1/k} + \text{HD})$ for the long-range scheme, and time $\mathcal{O}(|S_k| + \text{HD})$ for the additional broadcast step) and $\tilde{\mathcal{O}}(n^{1/(2k)})$, respectively, completing the proof. \blacksquare

We note that a distributed implementation of Thorup-Zwick distance oracles with stretch $2k-1$ and running time $\tilde{\Theta}(\text{SPD} \cdot n^{1/k})$ was recently given by Das Sarma et al. [6]. Intuitively, the reason for the discrepancy is that in [6], there is no use of the skeleton graph. In general, our running time and the one from [6] are incomparable (one may run both algorithms in parallel and use the output of the one that terminates first).

5.2 Approximate Weighted Diameter

Obtaining an approximation of the weighted diameter is simpler than constructing distance sketches. Dropping the short-range scheme from the construction, we can prove the following result.

Theorem 5.2 *For any $k \in \mathbb{N}$, the weighted diameter WD can be approximated w.h.p. to within a factor of $2k+1$ in the **CONGEST** model in $\tilde{\mathcal{O}}(n^{1/2+1/(2k)} + \text{HD})$ rounds.*

Proof: We use the following streamlined version of our algorithm.

1. Select a uniformly random skeleton S where $\Pr[v \in S] = 1/\sqrt{n}$ independently for all $v \in V$.
2. Apply [Algorithm 3](#) to construct a $(2k-1)$ -spanner of $G_{S,h(S_R)}$. Let WD^k be the weighted diameter of the spanner of $G_{S,h(S_R)}$ (which can be computed locally).
3. Apply Algorithm BSP, where all nodes in S function as the same source: $\text{source}(v) = 1$ for all $v \in S$ and $\text{source}(v) = \perp$ for all $v \notin S$. Use $\Delta = 1$ and $h \in \Theta(\sqrt{n} \log n)$.
4. Find the maximal distance d_{\max} computed by any node and output $2d_{\max} + \text{WD}^k$.

Regarding the time complexity, note that Step 2 requires $\tilde{\mathcal{O}}(n^{1/2+1/(2k)} + \text{HD})$ rounds by Theorem 4.17, Step 3 requires $\tilde{\mathcal{O}}(\sqrt{n})$ time by Theorem 4.4, and Step 4 takes $\mathcal{O}(\text{HD})$ rounds.

Regarding the approximation ratio, consider any $v, w \in V$, and let $s_v, s_w \in S$ the nodes in S closest to v and w , respectively. Then

$$\text{wd}(v, w) \leq \text{wd}(v, s_v) + \text{wd}(s_v, s_w) + \text{wd}(s_w, w) \leq d_{\max} + \text{WD}^k + d_{\max},$$

and hence $\text{WD} \leq \text{WD}^k + 2d_{\max}$. On the other hand, we have

$$\begin{aligned} \text{WD} &\geq \max \left\{ d_{\max}, \max_{s, t \in S} \{ W_{S, h(S_R)}(s, t) \} \right\} \\ &\geq \frac{2d_{\max} + (2k - 1) \max_{s, t \in S} \{ W_{S, h(S_R)}(s, t) \}}{2k + 1} \geq \frac{2d_{\max} + \text{WD}^k}{2k + 1}, \end{aligned}$$

where the last inequality holds w.h.p. by Theorem 4.17. \blacksquare

5.3 Distributed Approximation for Generalized Steiner Forest

In this section we explain how to utilize our routing scheme to obtain a fast distributed algorithm for the Generalized Steiner Forest problem (GSF), defined as follows.

Generalized Steiner Forest (GSF')

Input: A weighted graph $G = (V, E, W)$, a set of *terminals* $T \subseteq V$, and for each terminal $t \in T$ a component number $C(t)$.

Output: A subset of the edges $F \subseteq E$ such that for all pairs $s, t \in T$ with $C(s) = C(t)$, we have that s is connected to t in the graph (V, F) .

Goal: Minimize $\sum_{e \in F} W(e)$.

We note that sometimes, the connectivity requirement C is expressed as a set of node pairs. While the size of the input representation may differ, the two variants are equivalent for our purposes; using our spanner construction, we can obtain the component-based description within $\tilde{\mathcal{O}}(T + \text{HD})$ rounds from the pair-based formulation.

In the distributed setting, we assume that each node knows whether it is a terminal, and if so, what is its component number. Clearly, we can establish global knowledge on T and the component numbers within time $\mathcal{O}(T + \text{HD})$ by broadcasting the respective pairs of values over a BFS tree.

We now present a solution to GSF. We start with a generic reduction to a centralized algorithm which abstracts away the underlying graph G , and uses a graph whose nodes are just the terminals and edge weights are inter-terminal distance estimates.

To analyze Algorithm 5, we consider two simple transformations of the input instance and state their effect on the cost of the solution. First, consider the effect of using just distances between terminals (and not the whole graph). The following lemma bounds the effect of this simplification. Given an instance \mathcal{I} for GSF, we use $\text{OPT}(\mathcal{I})$ to denote any fixed optimal solution for \mathcal{I} .

Algorithm 5: Distributed algorithm for GSF. ALG is any centralized approximation algorithm for GSF'.

input : terminal components //locally known: v knows whether $v \in T$ and if so, its component number

output: F : edges in the Steiner forest

- 1 Obtain distance estimates for all distances $\text{wd}(s, t)$ with $s, t \in T$.
 - 2 Simulate ALG on the graph $G' = (T, E', W')$ and the same terminal components, where
 $E' := \{\{s, t\} \mid s, t \in T \wedge s \neq t\}$, and for all $s, t \in T$, $W'(s, t) := \min\{\tilde{\text{wd}}_s(t), \tilde{\text{wd}}_t(s)\}$ and $\tilde{\text{wd}}_s(t)$ denotes the estimate of $\text{wd}(s, t)$ computed at s (given the label of t). Denote by F' the computed solution.
 - 3 Identify and output all edges on paths in G that correspond to edges in F' .
-

Lemma 5.3 Let $\mathcal{I} = (G = (V, E, W), T, C)$ be an instance of GSF. Define an instance $\mathcal{I}' = (G', T, C)$, where $G' = (T, E', W')$ with $E' = \{\{s, t\} \mid s, t \in T\}$ and $W'(s, t) = \text{wd}_G(s, t)$. Then $W'(\text{OPT}(\mathcal{I}')) \leq 2W(\text{OPT}(\mathcal{I}))$.

Proof: The proof is a generalization of the standard argument for Steiner trees [28]. Let $F = \text{OPT}(\mathcal{I})$. Let $C_1 = (V_1, F_1), \dots, C_m = (V_m, F_m)$ be the connected components of F . By optimality of F , the components are trees. Fix a component C_j , and consider an Euler tour of its tree. Let $\sigma = \langle v_0, \dots, v_{2|V_j|} = v_0 \rangle$ be the sequence of nodes visited by the tour. Since each edge in F_j is visited exactly twice in σ , we have that the total weight of edges in the tour is $2W(F_j)$. Define the node sequence $\sigma' = \langle u_0, \dots, u_{|V_j|-1} \rangle$ obtained from σ by omitting the second occurrences of nodes from σ . Consider now the set of edges $F'_j := \{\{u_{i-1}, u_i\} \mid 0 < i < |V_j|\}$ in G' . Since the edges in G' are shortest paths in G , clearly the weight of F'_j is not more than the total weight of edges in σ , namely $W'(F'_j) \leq 2W(F_j)$. Finally, note that $F' = \bigcup_{j=1}^t F'_j$ is a feasible solution for \mathcal{I}' , and therefore

$$W'(\text{OPT}(\mathcal{I}')) \leq W'(F') = \sum_{j=1}^t W'(F'_j) \leq \sum_{j=1}^t 2W(F_j) = 2W(\text{OPT}(\mathcal{I})).$$

■

Next, consider replacing edge weights by ρ -approximate weights.

Lemma 5.4 Let $\mathcal{I} = (G = (V, E, W), T, C)$ and $\mathcal{I}' = (G = (V, E, W'), T, C)$ be instances of GSF differing only in the edge weights as follows: for all $e \in E$, $W(e) \leq W'(e) \leq \rho W(e)$ for some $\rho \geq 1$. Then $W(\text{OPT}(\mathcal{I}')) \leq \rho W(\text{OPT}(\mathcal{I}))$.

Proof: $W(\text{OPT}(\mathcal{I}')) \leq W'(\text{OPT}(\mathcal{I}')) \leq W'(\text{OPT}(\mathcal{I})) \leq \rho W(\text{OPT}(\mathcal{I}))$. ■

The preceding two lemmas show that using ρ -approximate distances and a centralized a -approximation algorithm for GSF, we will obtain a distributed $(2\rho a)$ -approximation algorithm for GSF.

It remains to show how to efficiently implement Algorithm 5 in the **CONGEST** model. The key is Step 1: Steps 2 and 3 will be performed locally at each node.

Corollary 5.5 For any integer $k \in [1, \log n]$, Algorithm 5 can be executed in the **CONGEST** model in $\tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ rounds with stretch factor $\rho(k) = 2k - 1$.

Proof: We apply the long-range routing scheme with skeleton set $S := T \cup R_S$, where R_S is sampled uniformly and independently at random with probability $n^{-1/2}$ from V . Lemma 4.15 implies that we

can perform Step 1 of Algorithm 5 within $\tilde{O}(n|S|^{1/k}/|S_R| + |S|^{1+1/k} + \text{HD}) = \tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ rounds with stretch $\rho(k) = 2k - 1$. Moreover, at the end of this step, all nodes know the spanner of the skeleton graph and can therefore locally compute W' . As remarked earlier, all nodes can learn the terminal components within $\tilde{O}(\sqrt{n} + \text{HD}) \subset \tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ rounds. With this information in place, all nodes can locally simulate ALG on G' and thus perform Step 2 of the algorithm. According to Lemma 4.16, the nodes on paths in G corresponding to edges in G' can learn of their membership within $\tilde{O}(n|S|^{1/k}/|S_R| + |S|^{1+1/k} + \text{HD})$ rounds as well. Afterwards, Step 3 of the algorithm can be completed locally as well. Summing up the running time bounds for the individual steps, we conclude that the overall time complexity is $\tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ as claimed. ■

Altogether, we arrive at the following result.

Theorem 5.6 *Given any integer $k \in [1, \log n]$ and any centralized a -approximation algorithm to GSF, GSF can be solved in the **CONGEST** model with approximation ratio $2a(2k-1)$ in $\tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ rounds, where T denotes the set of terminal nodes.*

Proof: Corollary 5.5 proves that Algorithm 5 can be implemented in $\tilde{O}((\sqrt{n} + |T|)^{1+1/k} + \text{HD})$ rounds, with distance estimates of stretch $\rho(k) = 2k - 1$. The approximation guarantee therefore follows from Lemmas 5.3 and 5.4. ■

We note that one can implement Step 1 also by computing distance sketches as in Theorem 4.19 without including the entire terminal set into the skeleton (i.e., G' does not become global knowledge), and simulate ALG sequentially, taking $\mathcal{O}(\text{HD})$ per step. This will reduce the overall running time in case $|T| \gg \sqrt{n}$ and HD times the step complexity of ALG (in terms of the number of globally synchronized steps) is small compared to $|T|^{1+1/k}$. However, this approach has two drawbacks. First, ALG cannot be arbitrary, but must admit to be simulated via a BFS tree using small messages. Second, the approximation ratio deteriorates according to the number of stages of the short-range scheme, as the number of stages contributes as a multiplicative factor ρ .

Discussion. It is known that the special case of MST, where all nodes are terminals in a single component has worst-case running time of $\tilde{\Omega}(\sqrt{n})$ even if the hop-diameter is $\mathcal{O}(\log n)$ [23]. However, it is unclear whether this lower bound holds if the number of terminals is small, and in turn, whether a larger number of terminal components makes the problem harder. For instance, for a single pair of terminals the problem reduces to selecting a single approximate shortest path; we are not aware of any non-trivial lower bound on this problem. Khan et al. [15] provide a $\mathcal{O}(\log n)$ -approximation to GSF within $\tilde{O}(\text{SPD} \cdot \gamma)$ rounds, where γ denotes the number of terminal components. The algorithm from [15] matches this bound up to factor $\log^{\mathcal{O}(1)} n$ if $\text{SPD} \cdot \gamma \in \tilde{O}(\sqrt{n})$; our approach does so in case $t \in \tilde{O}(\sqrt{n})$, where t is the number of terminal nodes. Note that the two running time bounds in general are incomparable: for approximation ratio $\mathcal{O}(\log n)$, we achieve time complexity $\tilde{O}(\sqrt{n} + t + \text{HD})$; there are instances for which $\text{SPD} \cdot \gamma \gg \sqrt{n} + t$ as well as those where $\sqrt{n} + t \gg \text{SPD} \cdot \gamma$. However, our approach is superior in that we can, for any integer $k \in [1, \log n]$, ensure an approximation ratio of $\mathcal{O}(k)$ at the expense of a slightly larger running time of $\tilde{O}((\sqrt{n} + t)^{1+1/k} + \text{HD})$ rounds. The authors of [15] employ probabilistic tree embeddings, a technique for which an approximation ratio of $\Omega(\log n)$ is inherent [10].

5.4 Tight Labels

The presented routing scheme relabels the nodes according to the Voronoi partition on each level. This yields suboptimal size of labels and makes it impossible for nodes to learn all labels $\lambda(V)$ quickly. We now present a modification of our routing scheme with labels $\lambda(V) = \{1, \dots, n\}$, trading in a larger stretch.

Instead of labeling the nodes on each level of the hierarchy independently, we do this by an inductive construction.

1. Define the partial order \prec on V given by $v \prec u$ if (and only if) one of the following is true:
 - $v, u \in S_L$ and the identifier of v is smaller than the identifier of u .
 - $l_v = l_u < L$, $Y_v(l_v + 1) = Y_u(l_v + 1)$, and v precedes u in a fixed DFS enumeration of the tree $T_{Y_v(l_v+1)}(l_v + 1)$ on $C_{Y_v(l_v+1)}(l_v + 1)$ induced by the shortest h_{l_v} -hop paths from each $w \in C_{Y_v(l_v+1)}(l_v + 1)$ to $Y_v(l_v + 1)$ detected by the invocation of Algorithm BSP in stage l_v .
2. Set $\text{count}_v(0) := 1$ for all $v \in V = S_0$. For each level $i \in \{1, \dots, L\}$, aggregate the sums of the values $\text{count}_s(i - 1)$ of nodes $s \in S_{i-1}$ in subtrees of $T_{Y_s(i)}(i)$ at the roots of these subtrees. We define for all $s \in S_i$ the value

$$\text{count}_s(i) := \sum_{\substack{s' \in S_{i-1} \\ s = Y_{s'}(i)}} \text{count}_{s'}(i - 1),$$

which can be computed from the received values. For each level i , this operation can be performed within $\mathcal{O}(h_i)$ rounds.

3. Each skeleton node $s \in S_L$ announces $\text{count}_s(L)$ to all other nodes. This requires $\mathcal{O}(|S_L| + \text{HD})$ rounds using a BFS tree.
4. Each skeleton node $s \in S_L$ sets

$$\lambda(s) := 1 + \sum_{\substack{s' \in S_L \\ s' \prec s}} \text{count}_{s'}(L).$$

5. Starting from level L and proceeding inductively on decreasing i , for each $i \in \{1, \dots, L\}$, each node $s \in S_i$, and each node $s' \in C_s(i) \cap S_{i-1}$, we inform s' of the value

$$\lambda(s') := \lambda(s) + 1 + \sum_{\substack{s'' \in C_s(i) \cap S_{i-1} \\ s'' \prec s'}} \text{count}_{s''}(i - 1).$$

Note that this step can be performed in $\mathcal{O}(h_i)$ rounds once λ_s is known due to the information collected in Step 2.

From the above arguments and the results from [Section 4.2](#) we can immediately conclude that the time complexity of computing these labels is negligible.

Corollary 5.7 *Executing the above construction does not increase the asymptotic time complexity of setting up the routing tables.*

By construction, $\lambda(V) = \{1, \dots, n\}$. Note that at the end of the above construction, each node v knows $\lambda(v)$ and, for each level i and each of its children in $T_{Y_v(i)}(i)$, it knows the range of labels

associated with this child. For each $v \in V$, set $v_0 := v$ and define inductively for $i \in \{1, \dots, L\}$ that $v_i = Y_{v_{i-1}}(i)$. Given any label $\lambda(w)$, we can thus route from any node v to w as follows.

1. Set $i := 0$.
2. If $i < L$ and $w_i \notin H_{v_i}(i+1)$, then route to v_{i+1} , set $i := i+1$, and repeat this step. If $i < L$ and $w_i \in H_{v_i}(i+1)$, route to w_i and proceed to the next step. If $i = L$, route to w_L using the long-range scheme and proceed to the next step.
3. If $i > 0$, then route to w_{i-1} , set $i := i-1$, and repeat this step. Otherwise $w_i = w$ and we are done.

The constructed sequence of routing indirections is thus $(v_0, \dots, v_{i_0}, w_{i_0}, \dots, w)$, where either i_0 is the minimal level such that $w_{i_0} \in H_{v_{i_0}}(i_0+1)$ or $i_0 = L$. As a result of these indirections, we cannot give a bound on the stretch that is linear in the number of levels anymore. However, we still can argue that $\text{wd}(v_i, w_i) \leq 4\text{wd}(v_{i-1}, w_{i-1})$ assuming that $w_{i-1} \notin H_{v_{i-1}}(i)$.

Lemma 5.8 *Suppose that for the labeling scheme stated above we have for some integer $1 \leq i_0 \leq L$ that $w_{i-1} \notin H_{v_i}(i)$ for all integers $0 \leq i < i_0$. Then*

$$\text{wd}(v_{i_0}, w_{i_0}) + \sum_{i=1}^{i_0} (\text{wd}(v_{i-1}, v_i) + \text{wd}(w_i, w_{i-1})) < 2 \cdot 4^{i_0} \text{wd}(v, w).$$

Proof: We show by induction that $\text{wd}(v_i, w_i) \leq 4^i \text{wd}(v, w)$ for all $0 \leq i \leq i_0$, which is obviously true for $i = 0$. Analogously to Lemma 4.6 we have that $\text{wd}(v_i, v_{i+1}) \leq \text{wd}(v_i, w_i)$ and $\text{wd}(w_i, w_{i+1}) \leq 2\text{wd}(v_i, w_i)$. By the triangle inequality,

$$\text{wd}(v_{i+1}, w_{i+1}) \leq \text{wd}(v_{i+1}, v_i) + \text{wd}(v_i, w_i) + \text{wd}(w_i, w_{i+1}) \leq 4\text{wd}(v_i, w_i) = 4^{i+1} \text{wd}(v, w).$$

This completes the induction and in addition reveals that

$$\sum_{i=1}^{i_0} (\text{wd}(v_{i-1}, v_i) + \text{wd}(w_i, w_{i-1})) \leq 3 \sum_{i=0}^{i_0-1} \text{wd}(v_i, w_i) \leq 3 \sum_{i=0}^{i_0-1} 4^i \text{wd}(v, w) < 4 \cdot 4^{i_0-1} \text{wd}(v, w).$$

Therefore

$$\text{wd}(v_{i_0}, w_{i_0}) + \sum_{i=1}^{i_0} (\text{wd}(v_{i-1}, v_i) + \text{wd}(w_i, w_{i-1})) < 2 \cdot 4^{i_0} \text{wd}(v, w),$$

concluding the proof. \blacksquare

As by Corollary 5.7 the construction time of the routing scheme is not affected by the above labeling and routing mechanism and Lemma 5.8 provides a stretch bound of $2 \cdot 2^{2L}$ for the modified short-range routes, we obtain the following statement.

Theorem 5.9 *Given $\alpha \in [1/2, 1]$, let $k = \lceil 1/(2\alpha - 1) \rceil$ if $\alpha > 1/2 + 1/\log n$ and $k = \lfloor 1/\log n \rfloor$ otherwise. Tables for stateless routing and distance approximation with stretch $\rho(\alpha) = 4k \cdot 4^{\lceil \log(k+1) \rceil} + 2k - 1 \in \mathcal{O}(k^3)$ with node labels $1, \dots, n$ can be constructed in the **CONGEST** model in $\tilde{\mathcal{O}}(n^\alpha + \text{HD})$ rounds.*

Acknowledgements

We would like to thank David Peleg for valuable discussions.

References

- [1] I. Abraham, C. Gavoille, and D. Malkhi. On Space-Stretch Trade-Offs: Lower Bounds. In *Proc. 18th Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pages 207–216, 2006.
- [2] I. Abraham, C. Gavoille, D. Malkhi, N. Nisan, and M. Thorup. Compact name-independent routing with minimum stretch. *ACM Transactions on Algorithms*, 4(3):37:1–37:12, 2008.
- [3] S. Baswana and S. Sen. A Simple and Linear Time Randomized Algorithm for Computing Sparse Spanners in Weighted Graphs. *Random Structures and Algorithms*, 30(4):532–563, 2007.
- [4] R. E. Bellman. On a routing problem. *Quart. Appl. Math.*, 16:87–90, 1958.
- [5] P. Chalermsook and J. Fakcharoenphol. Simple distributed algorithms for approximating minimum steiner trees. In *Computing and Combinatorics*, volume 3595 of *LNCS*, pages 380–389, 2005.
- [6] A. Das Sarma, M. Dinitz, and G. Pandurangan. Efficient Computation of Distance Sketches in Distributed Networks. In *Proc. 24th Symposium on Parallelism in Algorithms and Architectures*, 2012.
- [7] A. Das Sarma, S. Holzer, L. Kor, A. Korman, D. Nanongkai, G. Pandurangan, D. Peleg, and R. Wattenhofer. Distributed Verification and Hardness of Distributed Approximation. In *Proc. 43th ACM Symposium on Theory of Computing*, pages 363–372, 2011.
- [8] M. Elkin. An Unconditional Lower Bound on the Time-Approximation Tradeoff for the Minimum Spanning Tree Problem. *SIAM Journal on Computing*, 36(2):463–501, 2006.
- [9] M. Elkin and D. Peleg. $(1 + \epsilon, \beta)$ -Spanner Constructions for General Graphs. *SIAM Journal on Computing*, 33(3):608–631, 2004.
- [10] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proc. 35th Ann. Symp. on Theory of computing*, STOC ’03, pages 448–455, 2003.
- [11] L. R. Ford. Network flow theory. Technical Report P-923, The Rand Corp., 1956.
- [12] S. Frischknecht, S. Holzer, and R. Wattenhofer. Networks Cannot Compute their Diameter in Sublinear Time. In *Proc. 23rd Symposium on Discrete Algorithms*, pages 1150–1162, 2012.
- [13] J. A. Garay, S. Kutten, and D. Peleg. A Sub-Linear Time Distributed Algorithm for Minimum-Weight Spanning Trees. In *Proc. 34th Conference on Foundations of Computer Science (FOCS)*, pages 659–668, 1993.

- [14] S. Holzer and R. Wattenhofer. Optimal Distributed All Pairs Shortest Paths and Applications. In *Proc. 31st Symposium on Principles of Distributed Computing (PODC)*, July 2012.
- [15] M. Khan, F. Kuhn, D. Malkhi, G. Pandurangan, and K. Talwar. Efficient Distributed Approximation Algorithms via Probabilistic Tree Embeddings. *Distributed Computing*, 25:189–205, 2012.
- [16] M. Khan and G. Pandurangan. A Fast Distributed Algorithm for Minimum Spanning Trees. *Distributed Computing*, 20:391–402, 2008.
- [17] S. Kutten and D. Peleg. Fast Distributed Construction of Small k -Dominating Sets and Applications. *Journal of Algorithms*, 28(1):40–66, 1998.
- [18] Y. Mansour and B. Patt-Shamir. Greedy packet scheduling on shortest paths. *Journal of Algorithms*, 14:449–465, 1993.
- [19] J. McQuillan, I. Richer, and E. Rosen. The new routing algorithm for the arpanet. *IEEE Trans. on Communications*, COM-28(5):711–719, May 1980.
- [20] J. Moy. OSPF version 2. RFC 2328, Network Working Group, 1998.
- [21] D. Peleg. *Distributed Computing: A Locality-Sensitive Approach*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.
- [22] D. Peleg, L. Roditty, and E. Tal. Distributed Algorithms for Network Diameter and Girth. In *Proc. 39th Colloquium on Automata, Languages, and Programming (ICALP 2012)*. Springer Berlin / Heidelberg, 2012.
- [23] D. Peleg and V. Rubinovich. Near-tight Lower Bound on the Time Complexity of Distributed MST Construction. *SIAM Journal on Computing*, 30:1427–1442, 2000.
- [24] D. Peleg and E. Upfal. A trade-off between space and efficiency for routing tables. *Journal of the ACM*, 36(3):510–530, 1989.
- [25] Peleg, David and Schäffer, Alejandro A. Graph Spanners. *Journal of Graph Theory*, 13(1):99–116, 1989.
- [26] L. L. Peterson and B. S. Davie. *Computer Networks: A Systems Approach*. Morgan Kaufmann, 5th edition, 2011.
- [27] N. Santoro and R. Khatib. Labelling and Implicit Routing in Networks. *The Computer Journal*, 28:5–8, 1985.
- [28] H. Takahashi and A. Matsuyama. An approximate solution for the Steiner problem in graphs. *Math. Jap.*, 24:573–577, 1980.
- [29] M. Thorup and U. Zwick. Compact routing schemes. In *Proc. 13th Ann. ACM Symp. on Parallel Algorithms and Architectures*, SPAA ’01, pages 1–10, New York, NY, USA, 2001. ACM.
- [30] M. Thorup and U. Zwick. Approximate Distance Oracles. *Journal of the ACM*, 52(1):1–24, 2005.

- [31] U. Zwick. Exact and Approximate Distances in Graphs - A Survey. In *Proc. 9th European Symposium on Algorithms (ESA)*, pages 33–48, 2001.